October 2017

Asymdystopia: The threat of small biases in evaluations of education interventions that need to be powered to detect small impacts

John Deke

Thomas Wei

Tim Kautz

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE

Institute of Education Sciences
U.S. Department of Education

U.S. Department of Education
Betsy DeVos, *Secretary*

Institute of Education Sciences
Thomas W. Brock, *Commissioner for Education Research*
Delegated the Duties of Director

National Center for Education Evaluation and Regional Assistance
Ricky Takai, *Acting Commissioner*
Liz Eisner, *Acting Associate Commissioner*
Amy Johnson, *Project Officer*

NCEE 2018-4002

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

October 2017

This report was prepared for the Institute of Education Sciences (IES) under contract ED-IES-12-C-0083 by the Independent Review and Evaluation for Regional Educational Laboratories project administered by Mathematica Policy Research. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Deke, J., Wei, T., Kautz, T. (2017). *Asymdystopia: The threat of small biases in evaluations of education interventions that need to be powered to detect small impacts*. (NCEE 2018–4002). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

This report is available on the Institute of Education Sciences website at https://ies.ed.gov/ncee.

Abstract

Evaluators of education interventions are increasingly designing studies to detect impacts much smaller than the 0.20 standard deviations that Cohen (1988) characterized as "small." While the need to detect smaller impacts is based on compelling arguments that such impacts are substantively meaningful, the drive to detect smaller impacts may create a new challenge for researchers: the need to guard against smaller biases. The purpose of this paper is twofold. First, we examine the potential for small biases to increase the risk of making false inferences as studies are powered to detect smaller impacts, a phenomenon we refer to as asymdystopia. We examine this potential for two of the most rigorous designs commonly used in education research—randomized controlled trials (RCTs) and regression discontinuity designs (RDDs). Second, we recommend strategies researchers can use to avoid or mitigate these biases.

CONTENTS

Tables

Figures

Introduction

Evaluators of education interventions increasingly need to design studies to detect impacts much smaller than the 0.20 standard deviations that Cohen (1988) characterized as "small." For example, an evaluation of Response to Intervention from the Institute of Education Sciences (IES) detected impacts ranging from 0.13 to 0.17 standard deviations (Balu et al. 2015), and IES' evaluation of the Teacher Incentive Fund detected impacts of just 0.03 standard deviations (Chiang et al. 2015).

The drive to detect smaller impacts is in response to strong arguments that in many contexts, impacts once deemed "small" can still be meaningful (Kane 2015). Hill et al. (2008) and Lipsey et al. (2012) suggest multiple substantive benchmarks for assessing what a "meaningful" impact would be for a given intervention and context. These benchmarks often suggest that impacts less than 0.20 standard deviations are meaningful. For example, under the cost-effectiveness benchmark, smaller impacts may be deemed meaningful when evaluating less-expensive interventions.

Though based on a compelling rationale, the drive to detect smaller impacts may create a new challenge for researchers: the need to guard against relatively smaller biases. When studies were designed to detect impacts of 0.20 standard deviations or larger, it may have been reasonable for researchers to regard small biases as ignorable. For example, a bias of 0.03 standard deviations might have been ignorable in a study that could only detect an impact of 0.20 standard deviations. But in a study designed to detect much smaller impacts, such as Chiang et al. (2015) in which the impact estimate was 0.03 standard deviations, a bias of 0.03 standard deviations is no longer small—it is enormous.

The purpose of this paper is twofold. First, we examine the potential for small biases to increase the risk of making false inferences (in terms of the existence or magnitude of an impact) as studies are powered to detect smaller impacts. We refer to this phenomenon as asymdystopia.[1] We examine this potential for two of the most rigorous designs commonly used in education research—randomized controlled trials (RCTs) and regression discontinuity designs (RDDs). We focus on attrition bias in the case of RCTs and bias from regression misspecification in the case of RDDs. While the methodological details are distinct, in both cases we are unpacking a source of bias that may become increasingly problematic when studies are designed to detect smaller impacts. Second, we recommend strategies researchers can use to avoid or mitigate these biases.

More specifically, we address the following research questions:

1. **How problematic is attrition bias in RCTs as studies are powered to detect smaller impacts?** We explore this question using an attrition model for RCTs used in several federal evidence reviews. This model assumes that attrition bias is ignorable so long as it accounts for less than 20 percent of whatever size impact is deemed substantively important. Using this model and data on attrition from past studies, we examine:
    a. How attrition may become less acceptable, leading to higher rates of false inferences, as studies are powered to detect smaller effects;
    b. Contexts in which more favorable assumptions about the relationship among attrition, outcomes, and treatment status may allow for greater tolerance of attrition; and
    c. The feasibility of achieving lower attrition rates in future studies that are powered to detect small impacts, based on an analysis of attrition in past RCTs.

[1] Some studies—particularly retrospective nonexperimental studies using administrative data—have the statistical power to detect effects that are too small to be substantively important. This paper does not focus on "overpowered" studies. Instead, we focus on studies that are designed to have just enough statistical power to detect the smallest impact that is substantively important.

2. **How problematic is functional form misspecification bias in RDDs as studies are powered to detect smaller impacts?** In an RDD study, treatment and comparison groups are formed using a cutoff on a continuous assignment variable. When estimating RDD impacts, researchers must account for differences between the treatment and comparison group in the assignment variable. For example, a cutoff on a math test could be used to assign students to an intervention providing after-school homework help. In that example, students below the cutoff are in the treatment group and students above are in the comparison group. To estimate accurate impacts, researchers regression-adjust for the fact that students in the treatment group were lower math achievers to begin with. If the functional form for this regression is incorrect (for example, specifying a linear relationship when the true relationship is not linear), then the estimated impact could be biased. As a study's sample size increases, the bias due to functional form misspecification shrinks. At the same time, the precision of the estimates increases, which is typically a desirable property. However, a problem arises if the precision increases at a faster rate than the bias shrinks. In this situation, it is possible that rate of false inferences could increase, as researchers could find a statistically significant effect if the impact is precisely estimated but biased.

   We use Monte Carlo simulations to assess what happens as the sample size of the RDD increases under varying assumptions regarding the true functional form. Specifically, we examine the effect of a larger sample size on statistical power, functional form misspecification bias, and the accuracy of estimated *p*-values (or confidence intervals). We also assess whether a method proposed by Calonico et al. (2014) can be used to calculate accurate *p*-values, thereby reducing false inferences.

Overall, our findings suggest that biases that might have once been reasonably ignorable can pose a real threat in evaluations that are powered to detect small impacts. Our paper identifies and quantifies some of these biases and shows that they are important to consider when designing evaluations and when analyzing and interpreting evaluation findings. Our findings should *not* be interpreted as suggesting that researchers should avoid powering evaluations to detect small impacts. The problem of small biases is real but surmountable—so long as it is not ignored.

The remainder of the paper is organized into four sections. In the next section we discuss the motivation to detect smaller impacts and how this can lead to asymdystopia. The following two sections present the methods and findings corresponding to our two research questions. In the final section we conclude the paper with a discussion.

The power to detect small impacts and the potential for asymdystopia

Ideally, evaluations would be designed so that their minimum detectable effect (MDE) is calibrated to be the same as the smallest *substantively significant impact*. An impact is "detected" if it is *statistically significant*—that is, if the estimated impact is of a magnitude that it is very unlikely to occur when the true impact is zero. To detect smaller impacts with high probability, an evaluation typically needs a larger sample size. Because larger sample sizes lead to higher evaluation costs, researchers and funders typically seek to design studies that are just large enough to detect a substantively significant impact. See Murray (1998); Bloom (2004); Bloom et al. (2007); Hedges and Hedberg (2007); Schochet (2008a, 2008b); and Deke and Dragoset (2012) for more information about calculating statistical power in both RCTs and RDDs.

In his seminal book, Cohen (1988) suggested three thresholds researchers can use as a general guide for whether an impact is substantively significant or "meaningful." He suggested that impacts ranging from 0.20 to 0.49 are meaningful but "small." Impacts larger than 0.50 are "medium," and those exceeding 0.80 are "large." Cohen acknowledged that he based these thresholds on his own subjective judgments and advised caution in how they are applied. Still, the thresholds have been widely cited (Lipsey et al. 2012) and have served as benchmarks for some time in a number of fields, including education. For example, the What Works Clearinghouse (WWC) has long defined a "substantively important" impact to be *at least* 0.25 standard deviations (WWC 2008).[2] Similarly, many of IES' earlier evaluations were designed to detect impacts in the range of 0.20 to 0.25 (James-Burdumy et al. 2008, 2012; Agodini and Harris 2010).

Some researchers have argued more recently that using Cohen's benchmarks to design evaluations in education is often difficult to justify. Hill et al. (2008) and Lipsey et al. (2012) suggest a range of benchmarks for assessing what a "meaningful" impact would be for a given intervention in a given context. The benchmarks are:

1. **Normative expectations for academic growth**. This benchmark compares the impacts of an intervention to the growth in academic achievement that normally takes place over the course of one year. If an intervention is substantially less intensive than a full year of schooling, then we might expect it to have substantially smaller impacts than a year of schooling.
2. **Policy-relevant performance gaps**. This benchmark compares the impacts of an intervention to the difference in performance between two groups of students, for example, black and white students. If the purpose of an intervention is to substantially reduce this gap, then the current size of the gap is a relevant benchmark for choosing the evaluation's MDE and sample size.
3. **Observed impacts of similar interventions in similar contexts**. This benchmark examines the distribution of impact sizes on similar outcomes, for similar interventions, and in similar contexts to the evaluation being designed.
4. **Program impacts relative to cost.** This benchmark accounts for the cost of the intervention being evaluated relative to other interventions targeting similar outcomes in similar contexts. It might still be meaningful to detect a smaller impact for a less expensive intervention than was previously detected for more expensive interventions.

In addition, smaller impacts might be substantively meaningful for secondary outcomes that the intervention affects less directly. Many interventions are designed to have a large impact on a *proximal*

[2] The What Works Clearinghouse, managed by the U.S. Department of Education's Institute of Education Sciences, systematically reviews and synthesizes education research studies with the goal of providing a reliable source of scientific evidence for what works in education to improve student outcomes. For more information, see http://ies.ed.gov/ncee/wwc/.

outcome that is closely aligned to the intervention and often measured shortly after the end of the intervention. For example, a study of an after-school program offering help on homework might examine impacts on homework completion rates. Policymakers, however, might also be interested in *distal* outcomes that the intervention targeted less directly but could still be impacted. Continuing the example, improvements in homework completion might ultimately lead to gains on state achievement tests. The impact on distal outcomes is likely to be smaller than the impact on proximal outcomes because distal outcomes are influenced by a wider range of factors that are beyond the scope of the intervention to influence.

Perhaps reflecting these considerations, more recent education evaluations have sought to detect impacts on distal outcomes much smaller than 0.20 standard deviations. For example, IES' evaluation of the Teacher Incentive Fund had enough statistical power to detect impacts as small as 0.03 standard deviations (Chiang et al. 2015), while IES' evaluation of Response to Intervention was able to detect impacts ranging from 0.13 and 0.17 standard deviations (Balu et al. 2015). Requests for proposals to conduct new IES evaluations in the past few years have similarly asked offerors to detect impacts on student achievement from 0.10 to 0.15 (for example, the Impact Evaluation to Inform Teacher Preparation and Professional Development and the Impact Evaluation of Academic Language Interventions).

The potential for asymdystopia

Asymptopia has been described as a place where "data are unlimited and estimates are consistent" (Leamer 2010). An estimate is "consistent" if the expected value of the estimate approaches the true value of the parameter being estimated as the sample size approaches infinity. In other words, if we had unlimited data, we would get the right answer. Impact estimates from RCTs and RDDs are both consistent *if all the assumptions underpinning the methods are satisfied*.

Of course, asymptopia can never be achieved because data are never unlimited. That is, every study has a finite sample size. Nevertheless, it is tempting to believe that having more data is always a good thing. Specifically, it is tempting to believe that more data (1) always lead us closer to the correct answer and (2) always reduce the probability that we draw false inferences.

We define asymdystopia as a context in which a larger (but finite) sample size is not necessarily better and could even be worse from the perspective of controlling the Type 1 error rate. A Type 1 error occurs when an impact estimate is deemed statistically significant but the true impact is zero. The Type I error *rate* is the relative frequency of Type 1 errors across repeated impact estimates. For example, if the true impact of an intervention is zero, and we conduct 100 impact studies of the intervention, the Type 1 error rate is the expected proportion of those studies in which the impact is statistically significant. There has historically been a strong aversion to falsely concluding that an intervention works when in fact it does not. Researchers therefore typically prefer to limit the occurrence of Type 1 errors to 5 percent by only declaring an impact statistically significant if the *p*-value is 0.05 or less (or, equivalently, if the magnitude of t-statistic exceeds an appropriate cutoff). But if, as a study becomes larger, the standard error of the impact estimate shrinks while bias stays the same (or shrinks less than the standard error), then Type 1 errors could become more common. This is because the denominator of the t-statistic (the standard error) is shrinking faster than the numerator (the biased point estimate). For example, if the true impact is 0, bias is 0.05, and the standard error is 0.20, then the t-statistic is 0.05/0.20 = 0.25 (not statistically significant). If bias shrinks to 0.025 while the standard error shrinks to 0.01, then the t-statistic becomes 2.5 (statistically significant at the 5 percent level of significance).

How problematic is attrition bias in RCTs as studies are powered to detect smaller impacts?

Greenberg and Barnow (2014) identify sample attrition as potentially the most serious flaw that can lead to biased impact estimates in RCTs. Sample attrition occurs when individuals who were randomly assigned to treatment or control groups are missing outcome data for any reason. One key indicator of attrition bias is the attrition rate, both the overall rate for the study sample and the differential rates between treatment and control groups. Attrition bias is generally more concerning the larger the overall or differential attrition rates are. A second key indicator of attrition bias is how strongly attrition relates to outcomes and whether this relationship differs between treatment groups. For example, attrition bias would be high if outcome data were missing for the highest-achieving members of the control group and the lowest-achieving members of the treatment group. In general, the more strongly related these factors are, the more likely attrition bias is problematic. Unlike attrition rates, such relationships are unobservable, so some assumptions are needed. It is up to researchers to argue that the assumptions are plausible in their study's context.

To illustrate the problem posed by attrition bias in RCTs as studies are powered to detect smaller impacts, we first describe our model of attrition bias. Second, we examine how the tolerance level for overall and differential attrition changes as the target impact gets smaller. Third, we examine whether more favorable assumptions are needed about the relationship among attrition, outcomes, and treatment status as the target impact gets smaller. Finally, we examine the likely feasibility of designing studies powered to detect smaller impacts with attrition rates low enough to control bias at acceptable levels.

Summary of the WWC attrition model and standard[3]

We base our analysis on an existing attrition model developed by the WWC (2013, 2014). This model has been used to assess attrition bias in thousands of studies in education and other fields, making it a familiar model for many readers and well-suited to the types of analyses we conduct here.[4]

The model begins by assuming that all study participants have an unobserved latent propensity to stay in the study. The lower this propensity is, the more likely the study participant will attrite. This propensity, z, is assumed to be a normally distributed (0,1) random variable. If the total proportion of participants who stay in the study is denoted by P (and thus, the overall attrition rate is $1 - P$), and Φ is the standard normal cumulative distribution function, then participants will stay in the study if their z exceeds the threshold z^*, which is a deterministic function of P:

$$(1) \qquad z > \Phi^{-1}(1-P) = z^*$$

If y is the study outcome and also a normally distributed (0,1) random variable, then y is related to z as follows:

$$(2) \qquad y_t = \alpha_t z_t + u_t$$
$$y_c = \alpha_c z_c + u_c$$

Because this relationship may differ between treatment (t) and control (c) groups, there are two analogous equations subscripted by t and c. In this case, α is the correlation between z and y, whereas u is a normally distributed $\left(0, 1-\alpha^2\right)$ random variable independent of z. If α is 1 or –1, then all of y can be

[3] This summary draws heavily from the WWC's technical methods paper entitled *Assessing Attrition Bias* (http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=243), which includes complete details of the attrition model.

[4] The U.S. Department of Health and Human Services has also used this model. See, for example, the Home Visiting Evidence of Effectiveness Review (http://homvee.acf.hhs.gov) and the Teen Pregnancy Prevention Evidence Review (http://tppevidencereview.aspe.hhs.gov).

explained by *z*, whereas if α is zero, then *z* has no influence on *y*. Thus, the closer α is to zero, the less attrition is related to study outcomes, and, by extension, the less likely attrition would lead to biased impact estimates. The reverse is true as α gets closer to 1 or –1.

For simplicity, this model assumes that there are no impacts on mean outcomes in the study sample. Because there are no true impacts, an unbiased estimator should find no differences in expectation between treatment group outcomes and control group outcomes. Thus, attrition bias (*B*) is simply the expected difference between treatment group outcomes (y_t) and control group outcomes (y_c), which can be expressed using the following analytic formula, based on the properties of truncated normal distributions (ϕ is the standard normal probability density function):

$$
\begin{aligned}
(3) \qquad B &= E\left(y_t \mid z_t^*\right) - E\left(y_c \mid z_c > z_c^*\right) \\
&= \alpha_t E\left(z_t \mid z_t > z_t^*\right) - \alpha_c E\left(z_c \mid z_c > z_c^*\right) \\
&= \frac{\alpha_t \times \phi\left(\Phi^{-1}(1-P)\right)}{P_t} - \frac{\alpha_c \times \phi\left(\Phi^{-1}(1-P_c)\right)}{P_c}
\end{aligned}
$$

This result shows that attrition bias is driven by two main factors: the fraction of non-attriters (P_t and P_c), and the strength of the relationship between attrition and study outcomes (α_t and α_c). Moreover, the *differences* in these factors across treatment and control groups are important to consider. For example, if $P_t = P_c$ and $\alpha_t = \alpha_c$, there will be no attrition bias, even if a large proportion of the sample leaves and even if attrition is strongly related to outcomes. This result arises because the same types and fractions of participants drop out of both treatment and control groups. This uniformity preserves the equivalence of the remaining participants across both groups, leading to unbiased impact estimates. However, if either $P_t \neq P_c$ or $\alpha_t \neq \alpha_c$, then attrition bias will generally be present.

The analytic formula for attrition bias in Equation 3 allows us to precisely map out how much attrition bias exists for different combinations of attrition rates and α*'s*. The WWC uses two sets of assumptions for α. The conservative assumption sets $\alpha_t = 0.45$ and $\alpha_c = 0.39$. The optimistic assumption sets $\alpha_t = 0.27$ and $\alpha_c = 0.22$. The conservative and optimistic assumptions differ in two ways: (1) the degree to which study participants with outcome data differ from those without outcome data (that is, the size of α_t and α_c) and (2) the extent to which that relationship is itself related to treatment status (that is, how large the *difference* between α_t and α_c is). The optimistic assumption has a lower overall α_t and α_c, and a smaller difference between α_t and α_c. These assumptions imply that attrition is less related to the outcome and less related to treatment status, which suggests that all else equal, attrition bias would be less problematic.

It is not possible to estimate α_t and α_c directly. The WWC did, however, validate these parameter values based on empirical correlations between attrition and *baseline* measures of outcome variables, used as a proxy for the correlation between attrition and *follow-up* measures of those outcome variables. These correlations came from large-scale experimental evaluations of seven interventions (six curricular interventions and one teacher certification intervention) covering multiple grades and outcomes. They found that the observed correlations were generally most consistent with the optimistic assumption, but they retained the conservative assumption for special cases in which the treatment might plausibly have significant impacts on attrition.

For each of the two assumptions for α, it is possible to use Equation 3 to calculate the bias for various combinations of overall and differential attrition rates. More formally, the overall attrition rate is the proportion of randomized study participants who lack data on the evaluation's outcomes (equivalent to 1-*P* in Equation 3). The differential attrition rate is the difference between the treatment and control groups in the proportion of randomized study participants who lack data on the evaluation's outcomes (equivalent to $P_t - P_c$ in Equation 3). If the goal is to keep attrition bias within a certain maximum acceptable level, this exercise will reveal the acceptable combinations of overall and differential attrition rates. This method is exactly how the WWC derived its attrition standard.

The attrition standard aims to keep attrition bias to no more than 20 percent of the impact. Because the WWC defines a substantively important impact as 0.25 standard deviations, the maximum acceptable level of attrition bias is 0.05 standard deviations. By keeping attrition bias at this level, the Type 1 error rate is controlled at about 8 percent in studies that conduct hypothesis testing at the 5 percent significance level and that are powered to detect an impact of 0.25 standard deviations (with 80 percent power). In other words, the real Type 1 error rate is 8 percent compared to the nominal rate of 5 percent. We can calculate the real Type 1 error rate by calculating the power to detect an impact of the size of attrition bias. The power to detect an effect δ is given by Equation 4, where *T*(*x*,*df*,*ncp*) is the cumulative distribution function of the *t*-distribution evaluated at *x* with *df* degrees of freedom and a non-centrality parameter *ncp*, α is the probability of a Type 1 error, and n is the number of individuals randomized (this formula assumes that the treatment and control groups are of the same size). For example, a study that randomizes n=500 individuals has 80 percent power to detect an impact of $\delta = 0.25$ standard deviations, and there is an 8 percent chance of detecting an impact of 0.05 standard deviations (thus, when the only "impact" is due to bias, the Type 1 error rate is 8 percent).

$$(4) \qquad power(\delta) = 1 - T\left(T^{-1}(1-\alpha/2, df = n-2, ncp = 0), df = n-2, ncp = \frac{\delta}{\sqrt{4/n}} \right)$$

Figure 1 highlights the resulting bounds on overall and differential attrition rates. The green region shows combinations of overall and differential attrition rates that yield attrition bias less than or equal to 0.05 standard deviations under the conservative assumption. The yellow region shows combinations that yield acceptable bias under the optimistic assumption. The red region shows combinations that yield unacceptable bias under both sets of assumptions. Thus, in order to meet the WWC attrition standard, evaluators have tried to keep overall and differential attrition rates within the green or yellow regions.

Lower attrition rates might be needed to contain bias in studies powered to detect small impacts

Staying within the green and yellow regions in Figure 1 helps ensure that attrition bias is no larger than the maximum acceptable bias of 0.05 standard deviations. However, as studies are powered to detect impacts smaller than 0.25 standard deviations, the maximum acceptable bias also needs to be reduced accordingly to ensure that attrition bias accounts for no more than 20 percent of the impact and that the Type 1 error rate is controlled at an acceptable level. This means that if a study is powered to detect an impact of 0.10 standard deviations, attrition bias should be limited to 0.02 standard deviations. Similarly, if a study is powered to detect an impact of 0.05 standard deviations, attrition bias should be limited to 0.01 standard deviations. If the maximum acceptable bias is not reduced, then attrition bias could potentially account for most or all of the estimated impact leading to a much higher Type 1 error rate (Table 1), even if actual attrition levels fall within the green or yellow regions in Figure 1.

Table 1. The Type 1 error rate increases as studies are powered to detect smaller effects if attrition bias is held constant at 0.05 standard deviations

Magnitude of Impact Study Seeks to Detect	Type 1 Error Rate
0.25 (WWC definition of Substantively Important)	0.08
0.20	0.10
0.15	0.15
0.10	0.29
0.05	0.80

Source: Authors' calculations using equation 4.

Note: These calculations assume an RCT designed to detect a substantively important impact with 80 percent power at a significance level of 5 percent. The table shows that as studies are powered to detect smaller effects, the Type 1 error rate increases if attrition bias is held constant at 0.05 standard deviations.

To show how reducing the maximum acceptable bias from 0.05 to 0.02 or 0.01 affects attrition levels, we re-shade the green, yellow, and red regions of Figure 1. In Figure 2, we shade the areas that, based on Equation 3, yield bias of no more than 0.02 standard deviations (instead of 0.05). In Figure 3, we shade the areas that yield bias of no more than 0.01 standard deviations. The results in Figures 2 and 3 show that substantially tighter attrition bounds are needed. For example, assuming (1) the WWC's optimistic parameters, (2) no differences in attrition rates between treatment and control groups, and (3) a maximum acceptable bias of 0.05, the highest acceptable overall attrition rate is about 60 percent (Figure 1). All else equal, if the maximum acceptable bias is 0.02 instead of 0.05, then the analogous highest acceptable overall attrition rate drops from 60 percent to about 20 percent (Figure 2). If the maximum acceptable bias is 0.01, then the highest acceptable overall attrition rate drops to about 10 percent (Figure 3).

Figure 1. WWC Attrition Bounds

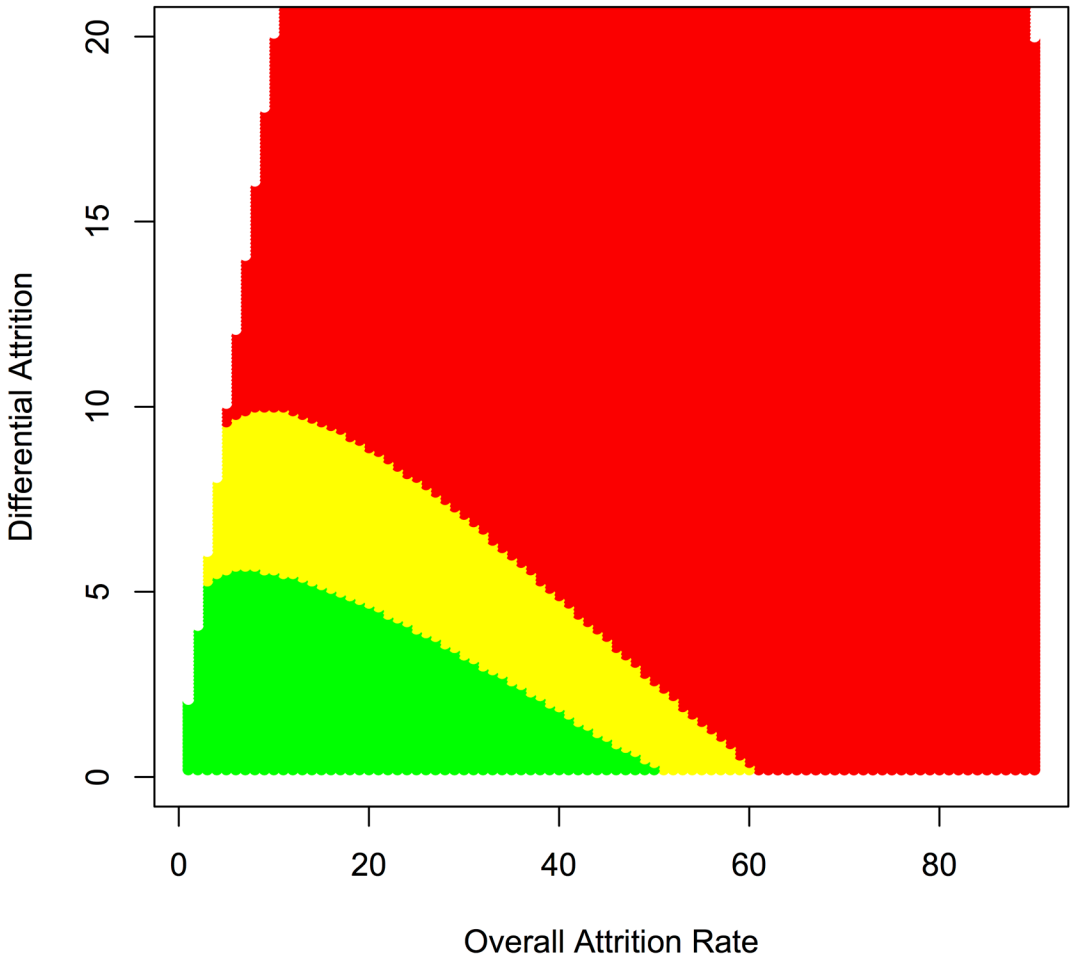


Figure 2. Attrition Bounds If the Maximum Acceptable Bias is 0.02 Standard Deviations

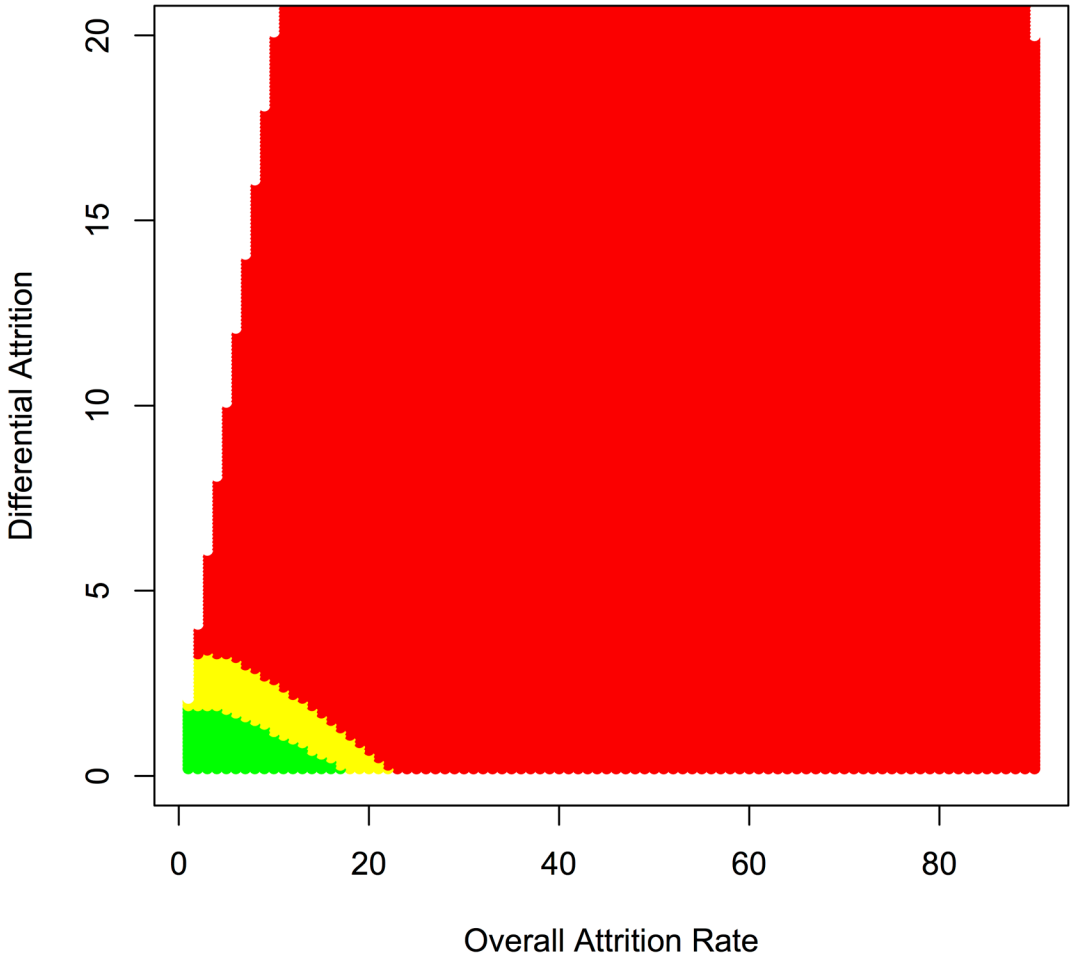
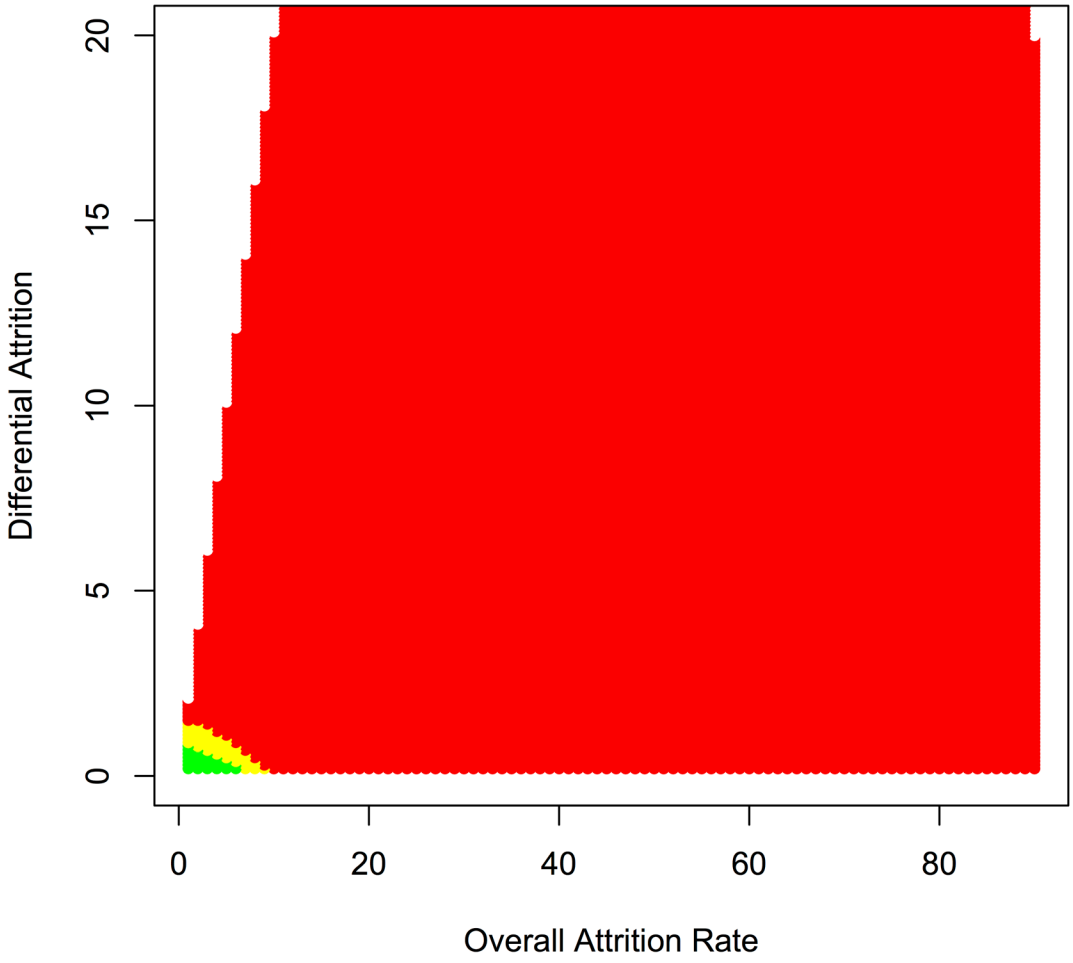


Figure 3. Attrition Bounds If the Maximum Acceptable Bias is 0.01 Standard Deviations



The highest acceptable differential attrition rate is also substantially smaller when limiting the maximum acceptable bias to 0.02 or 0.01. For example, in Table 2 we calculate the highest acceptable differential attrition rate when the overall rate is half of the maximum acceptable overall rate (the maximum acceptable overall attrition rates were presented in the previous paragraph). Under the WWC's optimistic assumptions, an overall attrition rate of 30 percent, and a maximum acceptable bias of 0.05, the highest acceptable differential attrition rate is about 6 percentage points (for example where treatment group attrition rate is 33% and control group attrition rate is 27%). If the maximum acceptable bias is 0.02 instead of 0.05, then the highest acceptable differential attrition rate is 2 percentage points, rather than 6 percentage points. If the maximum acceptable bias is 0.01, then the highest acceptable differential attrition rate is about 1 percent.

Table 2. Highest acceptable differential attrition rate

Highest acceptable bias (standard deviations)	Half of highest acceptable overall attrition rate (percent)	Highest acceptable differential attrition rate (percentage points)
0.05	30	6
0.02	10	2
0.01	5	1

Source: Authors' calculations.

Note: Highest acceptable overall attrition rate is the highest level of attrition at which bias is below the highest acceptable level given zero differential attrition.

The numeric examples in the previous two paragraphs are specific examples to illustrate more concretely how the tolerance levels for overall and differential attrition change as the maximum acceptable bias for attrition falls. Visually comparing Figures 1, 2, and 3 provides a more general picture of how the attrition bounds become substantially tighter overall.

Lower attrition rates might not be needed in some study contexts

The previous section's results show that *all else equal*, substantially lower levels of overall and differential attrition are needed to contain bias in studies powered to detect small impacts. Although researchers should strive to meet these targets, they may not always succeed in practice. Is all hope lost in these cases? Not necessarily, as Equation 3 shows that bias depends not only on the fraction of non-attriters (P_t and P_c) but also on how attrition, outcomes, and treatment status relate to one another (α_t and α_c). The previous section's analysis applied the standard WWC optimistic assumptions about this relationship. However, more favorable assumptions may be justifiable in some studies. If so, bias could still be contained to an acceptable level in these studies *even if the overall and differential attrition levels are in the standard WWC ranges* (Figure 1).

In this section, we examine just how much more favorable these assumptions would need to be for the standard WWC attrition bounds to be appropriate for studies powered to detect small impacts. To do so, we use Equation 3 to compute which values of α_t and α_c will contain bias to the lower levels needed (that is, 0.01 or 0.02 standard deviations, instead of the usual 0.05) in studies powered to detect small impacts, assuming that attrition levels fall within the typical bounds in Figure 1.[5] Table 3 reports the results for one set of attrition rates, but the basic conclusion that α_t and α_c would need to be more favorable holds more generally across all combinations of attrition rates. Recall that smaller overall α's and smaller differences between α_t and α_c are more favorable because they imply that attrition is less related to outcomes and treatment status and therefore less likely to bias estimated impacts. The results clearly show that as the maximum acceptable attrition bias falls for studies powered to detect small impacts, the model assumptions need to become more favorable for any given level of overall and differential attrition.

[5] For any observed overall and differential attrition rates, there are many values of α_t and α_c that would yield a given level of bias (see Equation 3). To calculate a unique pair of model parameters for each given level of bias, we assume that $\alpha_t = r\alpha_c$, where r is a constant equal to the ratio of α_t to α_c implicit in the WWC parameters (0.27/0.22). This approach allows us to uniquely characterize how optimistic the study parameters would need to be to contain bias.

Table 3. Assumptions needed to apply current attrition bounds with lower bias

	Acceptable attrition		Attrition model parameter assumptions	
Maximum acceptable bias	Overall	Differential	α_t	α_c
0.05	30	6	0.27	0.22
0.02	30	6	0.12	0.10
0.01	30	6	0.06	0.05

Source: Authors' calculations using attrition model described in Equation 3.

Note: The first row corresponds to the existing WWC optimistic attrition standard, which seeks to contain bias to 0.05 standard deviations. The second and third rows show how attrition model parameter assumptions would need to change to limit bias to 0.02 and 0.01 standard deviations at the same levels of overall and differential attrition. Values of α_t and α_c are correlations, the attrition rates are percentage points, and the maximum acceptable bias is standard deviation units.

Just how much more favorable are these assumptions? As noted earlier, the WWC previously calculated the correlations between attrition and baseline measures, used as a proxy for the correlation between attrition and outcome measures. Across seven interventions, they found that the correlation between baseline measures and attrition ranged from 0.01 to 0.28 for treatment groups and from 0.06 to 0.26 for control groups. Moreover, the treatment–control difference in correlations ranged from 0.01 to 0.10. Using these benchmarks, we see that the required assumptions calculated in Table 3 for lower levels of attrition bias (0.01 and 0.02) are within the empirically observed ranges, although they are at the more optimistic end of that range.

To gain an even better understanding of how much more optimistic these assumptions are, we simulated outcome and attrition data using the attrition rates and values of α_t and α_c shown in Table 3. These data were generated using the formulas in Equations 1 and 2, which means that the outcomes for the full sample (including both attriters and non-attriters) follow the standard normal distribution (mean zero, variance one).[6] We then calculated descriptive statistics for the attrited members of these simulated data sets as well as comparative density plots. Table 4 shows the descriptive statistics for three different scenarios. For scenario 1, we generated data with optimistic WWC values of α_t and α_c and attrition rates for the treatment and control groups that yield bias of 0.05 standard deviations. We report the mean of the outcome variable for the attrited sample in the treatment and control groups. We also report quartiles of the same variable. Scenarios 2 and 3 hold the attrition rates constant but change values of α_t and α_c to yield biases of 0.02 and 0.01 standard deviations.

[6] Note that it does not matter which attrition rates and values of α correspond to the treatment or control groups—switching all treatment and control labels would still yield the same conclusions.

Table 4. Outcomes of attrited and non-attrited samples generated under varying assumptions

		Outcomes of attriters and non-attriters		
α	Attrition rate	Mean of attriters	Mean of non-attriters	Difference in means (mean of non-attriters – mean of attriters)
Scenario 1: Attrition bias of 0.05 under WWC optimistic parameter assumptions				
α_t =0.27	33	-0.30	0.15	0.44
α_c =0.22	27	-0.27	0.10	0.37
Scenario 2: Parameter assumptions that yield attrition bias of 0.02 under scenario 1 attrition rates				
α_t =0.12	33	-0.13	0.06	0.20
α_c =0.10	27	-0.12	0.05	0.17
Scenario 3: Parameter assumptions that yield attrition bias of 0.01 under scenario 1 attrition rates				
α_t =0.06	33	-0.07	0.03	0.10
α_c =0.05	27	-0.06	0.02	0.08

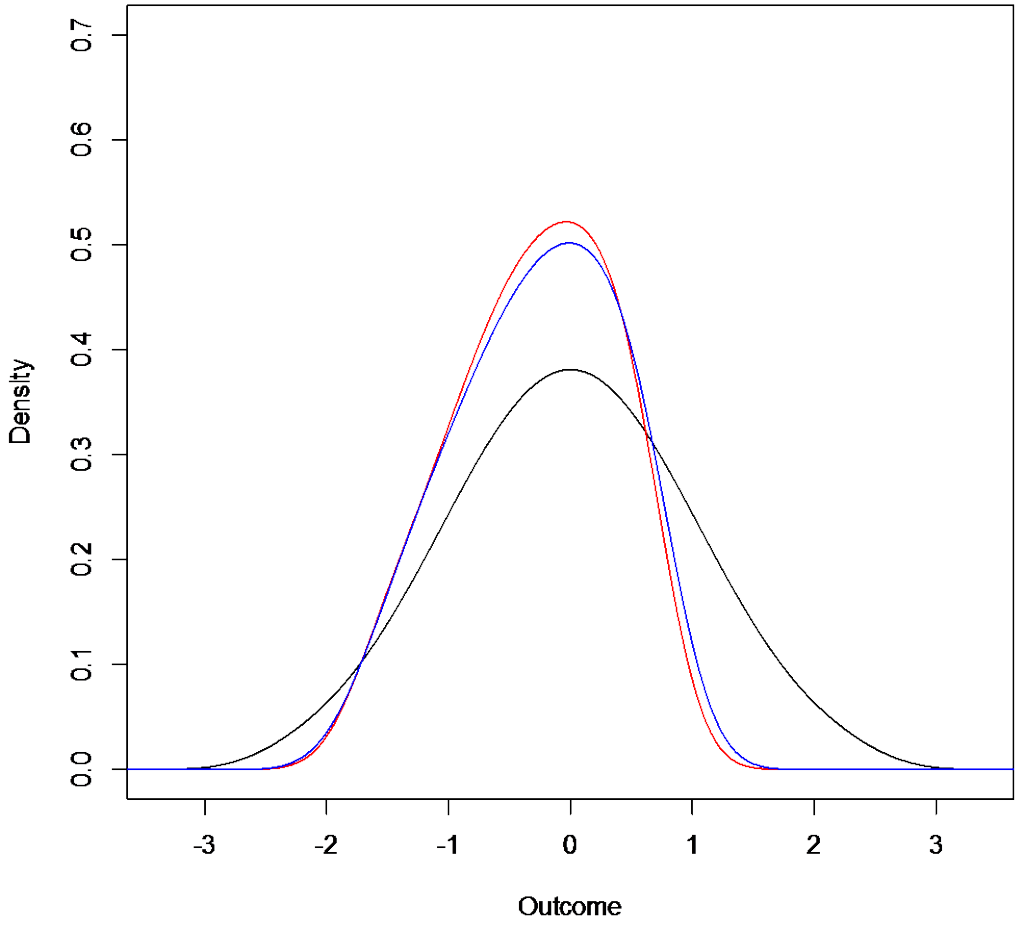
Source: Authors' calculations using attrition model described in Equation 3.

Note: The first row for each scenario is the treatment group, the second row is the control group. Values of α_t and α_c are correlations, the attrition rates are percentage points, and the descriptive statistics are standard deviation units.

The descriptive statistics in Table 4 show that to apply the existing WWC attrition bounds for lower levels of acceptable bias, we must effectively assume that the participants who leave a study's sample are very similar to those who stay, and that the participants who leave the treatment group are very similar to those who leave the control group. First, there is a much smaller difference in outcomes between participants who leave the study and those who stay. Under the WWC optimistic assumptions (scenario 1), follow-up test scores of participants who leave the study are about 0.44 to 0.37 standard deviations lower than those of participants who stay. But under the assumptions needed to limit bias to 0.02 or 0.01 standard deviations (scenarios 2 and 3), this gap must fall to as little as about 0.10 to 0.08 standard deviations. Second, Table 4 shows a smaller difference between the treatment and control groups in the characteristics of the attrited sample (meaning that the intervention had a smaller differential impact on the types of participants who left the treatment group versus the control group). This difference was already assumed to be modest under WWC assumptions (0.03 standard deviations under scenario 1, the difference between -0.30 and -0.27), but it becomes even smaller (0.01 standard deviations) under the more favored scenarios needed for studies powered to detect smaller impacts (scenarios 2 and 3).

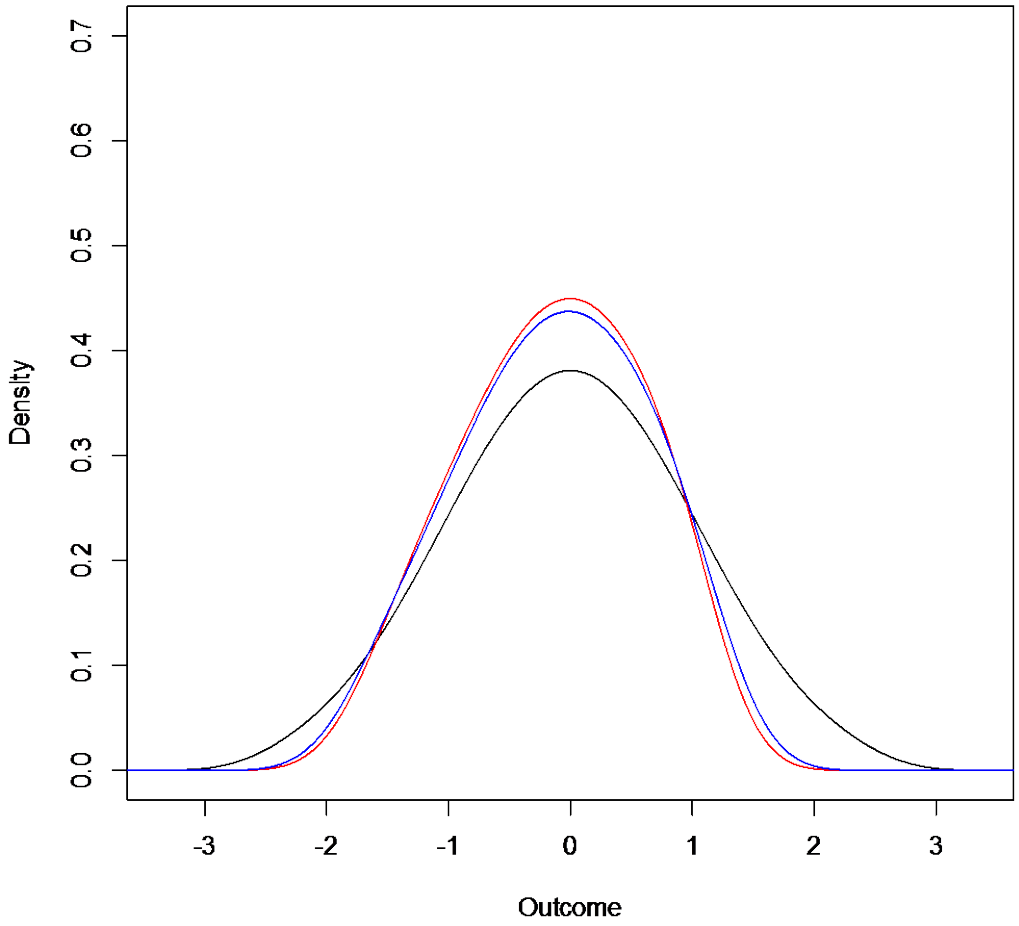
We further illustrate in Figures 4–6 the characteristics of the attrited samples under each of the three scenarios presented in Tables 3 and 4. Using kernel density plots, these figures show the outcome distribution for three samples: (1) the full study sample (shown in black), (2) the attrited sample from the control group (shown in blue), and (3) the attrited sample from the treatment group (shown in red). The figures reinforce the findings of Table 4—as the maximum acceptable bias is reduced in studies powered to detect smaller impacts, more optimistic assumptions about the relationship among attrition, outcomes, and treatment are needed. In particular, as the maximum acceptable bias falls in studies powered to detect smaller impacts, the attrited samples from the treatment and control groups need to become more similar to each other and more similar to the overall sample.

Figure 4. Comparative density plots for the sample described in Table 4 Scenario 1



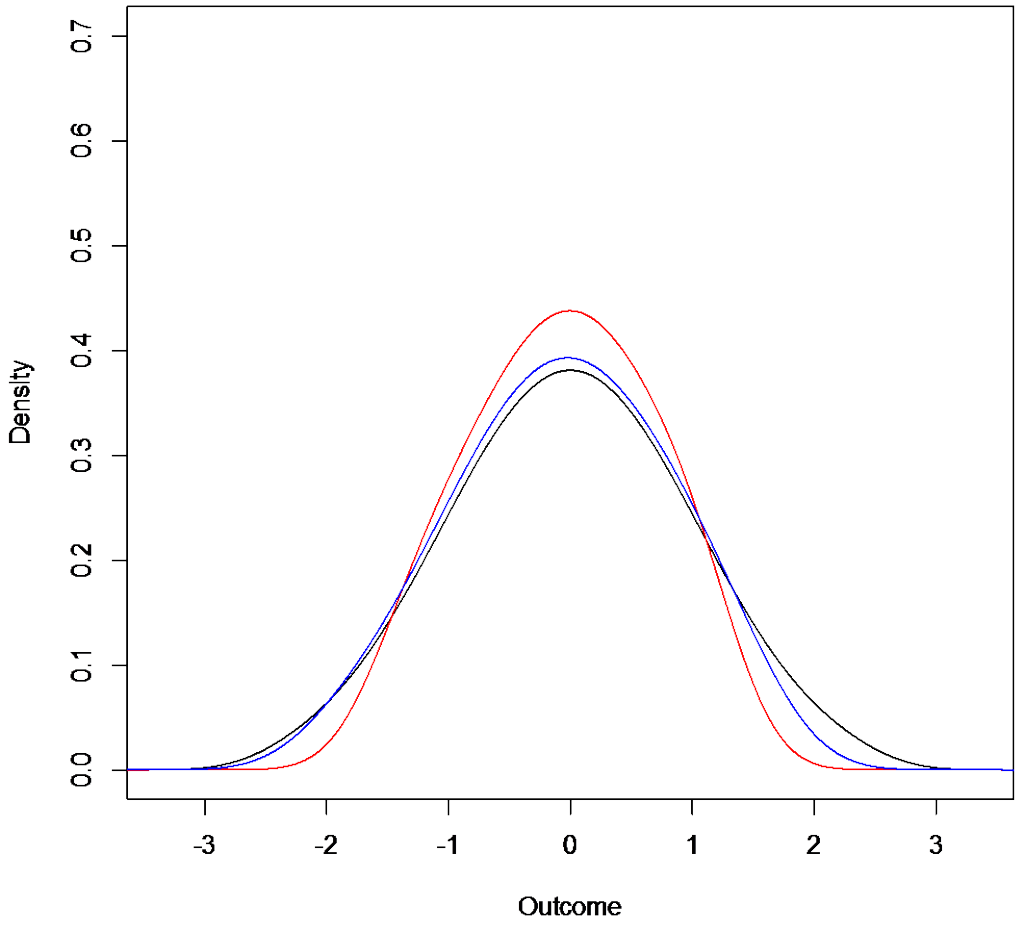
Note: The black density plot corresponds to the full sample. The blue density plot corresponds to attrited students from the comparison group. The red density plot corresponds to attrited students from the treatment group. Scenario 1 is based on the following assumptions: $\alpha_t = 0.27$, $\alpha_c = 0.22$, the attrition rate in the treatment group is 33 percent, and the attrition rate in the comparison group is 27 percent.

Figure 5. Comparative density plots for the sample described in Table 4 Scenario 2



Note: The black density plot corresponds to the full sample. The blue density plot corresponds to attrited students from the comparison group. The red density plot corresponds to attrited students from the treatment group. Scenario 2 is based on the following assumptions: $\alpha_t = 0.12$, $\alpha_c = 0.10$, the attrition rate in the treatment group is 33 percent, and the attrition rate in the comparison group is 27 percent.

Figure 6. Comparative density plots for the sample described in Table 4 Scenario 3



Note: The black density plot corresponds to the full sample. The blue density plot corresponds to attrited students from the comparison group. The red density plot corresponds to attrited students from the treatment group. Scenario 3 is based on the following assumptions: $\alpha_t = 0.06$, $\alpha_c = 0.05$, the attrition rate in the treatment group is 33 percent, and the attrition rate in the comparison group is 27 percent.

Data from past studies show that attaining lower attrition rates is difficult, but not impossible

When researchers design their studies to detect smaller impacts and still want to ensure that attrition bias accounts for no more than 20 percent of their smallest detectable impact, they need to consider whether they can realistically achieve lower attrition rates. To investigate whether lower attrition is feasible in practice, we used the study review database from the WWC to examine how often past studies achieved overall and differential attrition rates consistent with limiting bias to no more than 0.02 or 0.01 standard deviations (corresponding to study MDEs of 0.10 or 0.05 standard deviations).

From the WWC database we focused on RCTs that received a rating of *Meets WWC Standards Without Reservations* because these represent well-executed studies that provide a natural benchmark for considering the feasibility of achieving lower levels of attrition.[7] We supplemented the WWC downloadable database with additional information on sample sizes and attrition rates from the WWC master review guides. We only included analyses that affected whether each RCT met WWC standards,

[7] We exclude quick reviews because the review protocol differs from other types of reviews. The database is available at https://ies.ed.gov/ncee/wwc/StudyFindings.

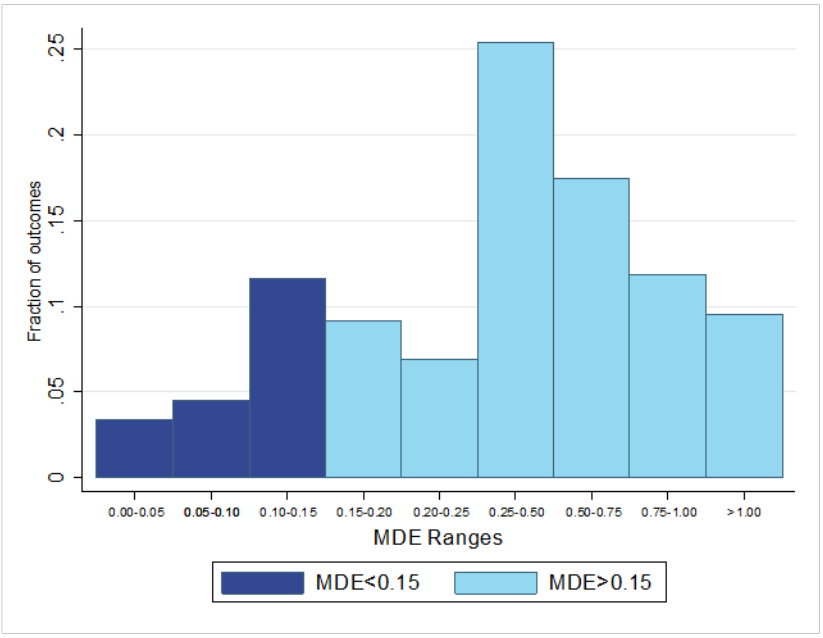
and thereby excluded supplementary analyses such as the impacts on subgroups. For each study,[8] we calculated the MDE using the *p*-value, effect size, and analytical sample size.[9]

Studies with low MDEs (which we define to mean less than 0.15 standard deviations) represent approximately 20 percent of all studies and have similar characteristics to those with high MDEs. Figure 7 shows the distribution of estimated MDEs across studies. Table 5 provides descriptive statistics for the full sample, as well as subsamples based on whether the MDE is less than or greater than 0.15. Overall, the studies with low MDEs have similar characteristics to those with high MDEs, with the exception that studies with low MDEs are more likely to focus on older students and have lower impacts.

[8] We use the term "study" to refer to outcome-intervention combinations because different outcomes can have different MDEs within a given evaluation.

[9] For each study, we calculate the MDE using the following formula, $MDE = \left[T^{-1}\left(N-1, 1-\frac{\alpha}{2}\right) + T^{-1}(N-1, \beta) \right] \left| ES / T^{-1}\left(N-1, \frac{P}{2}\right) \right|$, where T^{-1} is the inverse *t*-distribution, α is the significance level (assumed to be 0.05), β is the power (assumed to be 0.80), *ES* is the effect size, *p* is the *p*-value, and *N* is the analytical sample size for the unit of randomization.

Figure 7. Distribution of estimated minimum detectable effect sizes across studies



Source: Authors' calculations using WWC database.

Table 5. WWC Study sample characteristics

Variable	Sample		
	Full sample	MDE < 0.15	MDE > 0.15
Total number of studies	869	170	699
Distribution of overall attrition rates			
Mean	0.13	0.14	0.12
25th percentile	0.00	0.01	0.00
Median	0.09	0.15	0.09
75th percentile	0.20	0.20	0.19
Distribution of differential attrition rates			
Mean	0.02	0.02	0.03
25th percentile	0.00	0.00	0.00
Median	0.01	0.01	0.01
75th percentile	0.04	0.02	0.04
Distribution of impacts (effect size units, absolute value)			
Mean	0.27	0.05	0.32
25th percentile	0.06	0.02	0.08
Median	0.13	0.05	0.19
75th percentile	0.35	0.08	0.45
Percentage of studies from clustered designs (vs. non-clustered)	24.9	25.3	24.7
Percentage of studies by target population			
Elementary school students (or below)	51.6	30.6	56.9
Middle school students	16.3	10.0	17.8
High school students (or above)	32.1	59.4	25.3
Percentage of studies by type of intervention			
Practice	19.0	13.5	20.3
Supplement	2.6	5.9	1.9
Teacher-level	18.6	17.6	18.9
Curriculum	8.9	14.7	7.4
Policy	43.0	39.4	43.9
School-level	7.8	8.8	7.6

Source: Authors' calculations using WWC database and supplemental attrition data from WWC master review guides.

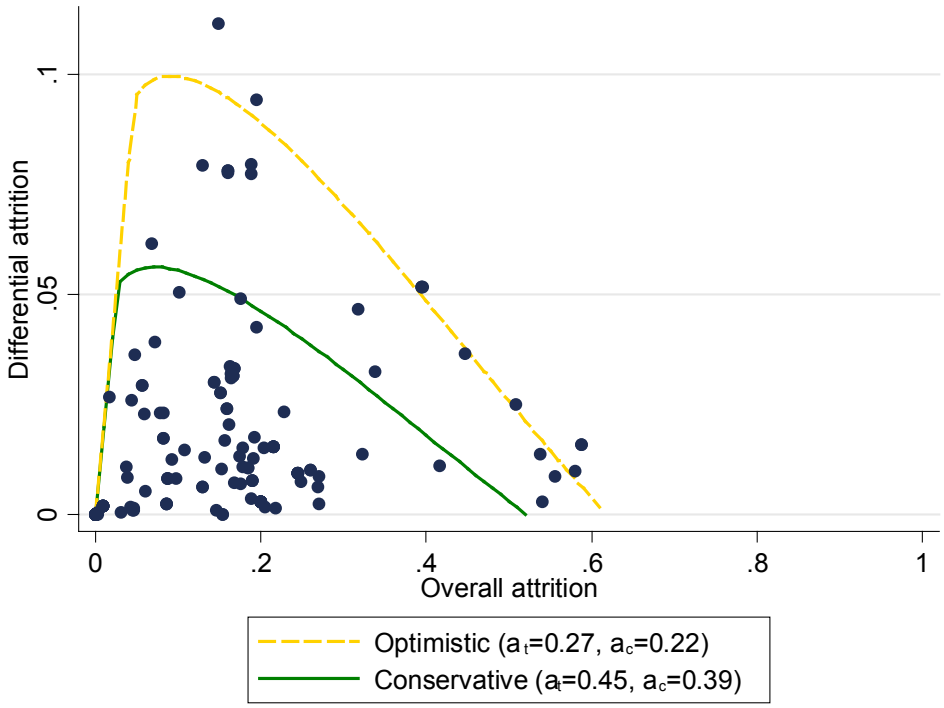
Under the WWC's optimistic parameter assumptions, over half of studies with low MDEs have attrition rates low enough to keep bias below 0.02 standard deviations and one third have attrition rates low enough to keep bias below 0.01 standard deviations (Table 6, row 3). In study contexts where even more optimistic assumptions are appropriate, these percentages can be much higher. With a bias threshold of 0.02, 92 percent of studies have acceptable attrition under the more optimistic parameters considered earlier, $\alpha_t = 0.12$ and $\alpha_c = 0.10$ (Table 6, row 2, column 2). With a bias threshold of 0.01, 92 percent of studies have acceptable attrition under the most optimistic parameters considered earlier, $\alpha_t = 0.06$ and $\alpha_c = 0.05$ (Table 6, row 1, column 1). We also present this information graphically in Figures 8-10. These figures are similar in construction to Figure 1 in that they show the combinations of overall and differential attrition needed to keep attrition bias below a specified level under varying assumptions regarding attrition model parameters. In Figure 8, the maximum acceptable bias is 0.05 standard deviations, in Figure 9 it is 0.02 standard deviations, and in Figure 10 it is 0.01 standard deviations.

Table 6. The percentage of past studies with acceptable attrition under three different maximum acceptable bias thresholds and three attrition model parameters assumptions

Attrition model parameters		Percentage of past studies with acceptable attrition under three maximum acceptable bias thresholds		
α_t	α_c	0.01	0.02	0.05
0.06	0.05	92	100	100
0.12	0.10	61	92	100
0.27	0.22	33	57	95

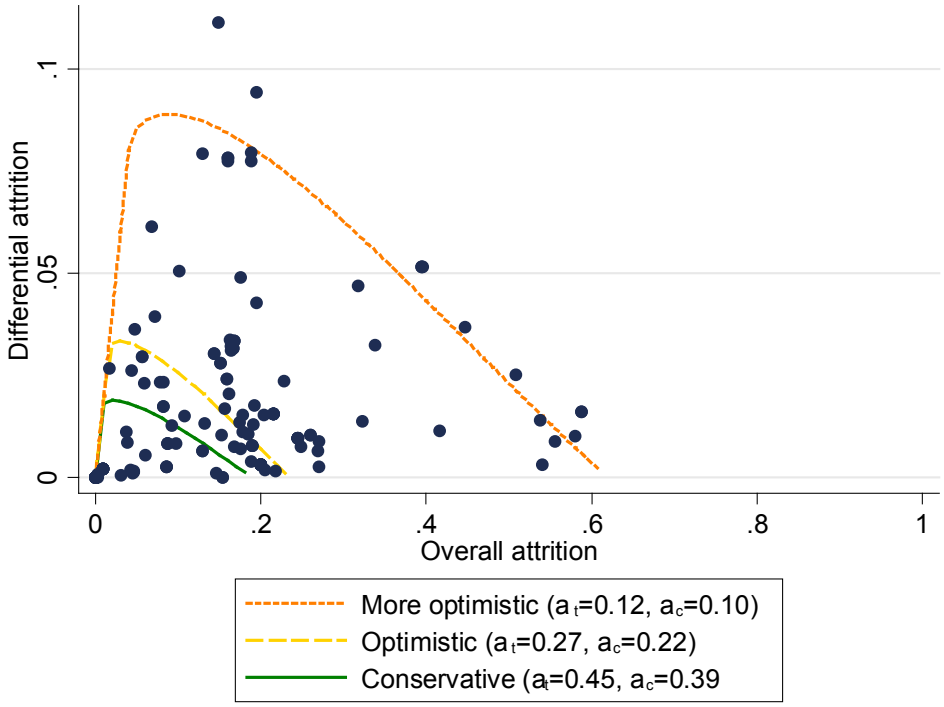
Source: Authors' calculations using WWC database and supplemental attrition data from WWC master review guides.

Figure 8. Among designs with MDEs <0.15, overall and differential attrition rates and attrition bounds if the maximum acceptable bias is 0.05 standard deviations



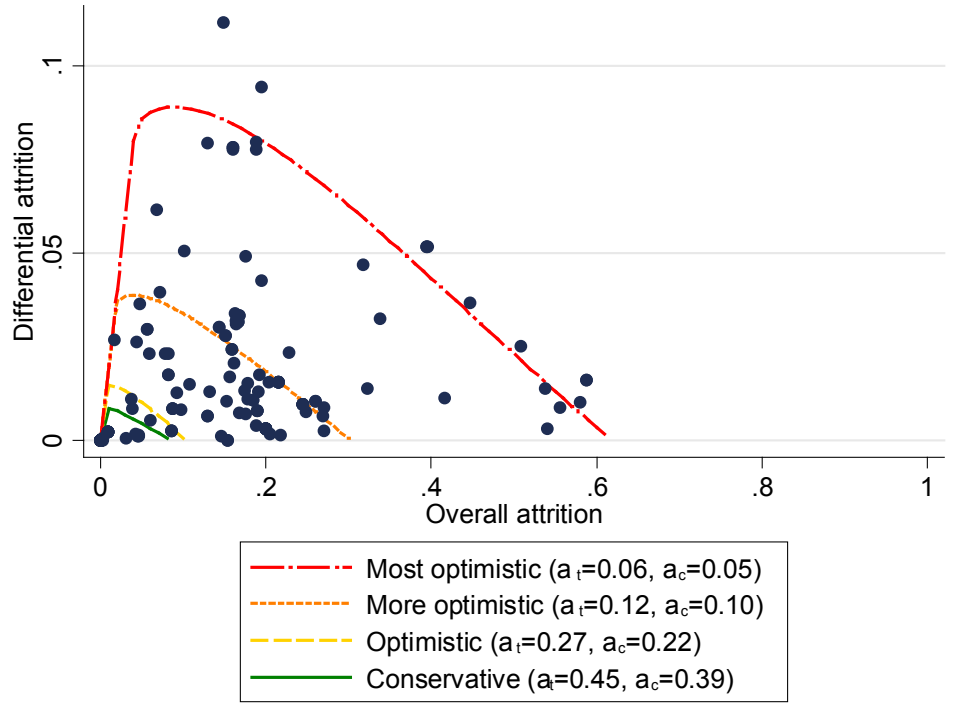
Source: Authors' calculations using WWC database and supplemental attrition data from WWC master review guides.

Figure 9. Among designs with MDEs <0.15, overall and differential attrition rates and attrition bounds if the maximum acceptable bias is 0.02 standard deviations



Source: Authors' calculations using WWC database and supplemental attrition data from WWC master review guides.

Figure 10. Among designs with MDEs <0.15, overall and differential attrition rates and attrition bounds if the maximum acceptable bias is 0.01 standard deviations



Source: Authors' calculations using WWC database and supplemental attrition data from WWC master review guides.

Researchers need to carefully consider whether their study context warrants more optimistic parameter assumptions. If more optimistic assumptions are made when they are unwarranted, the result could be a low-quality study with misleading findings. Recall that attrition is particularly problematic when students with missing data in the treatment group are fundamentally different from students with missing data in the control group. There are several scenarios where this is possible, including the following:

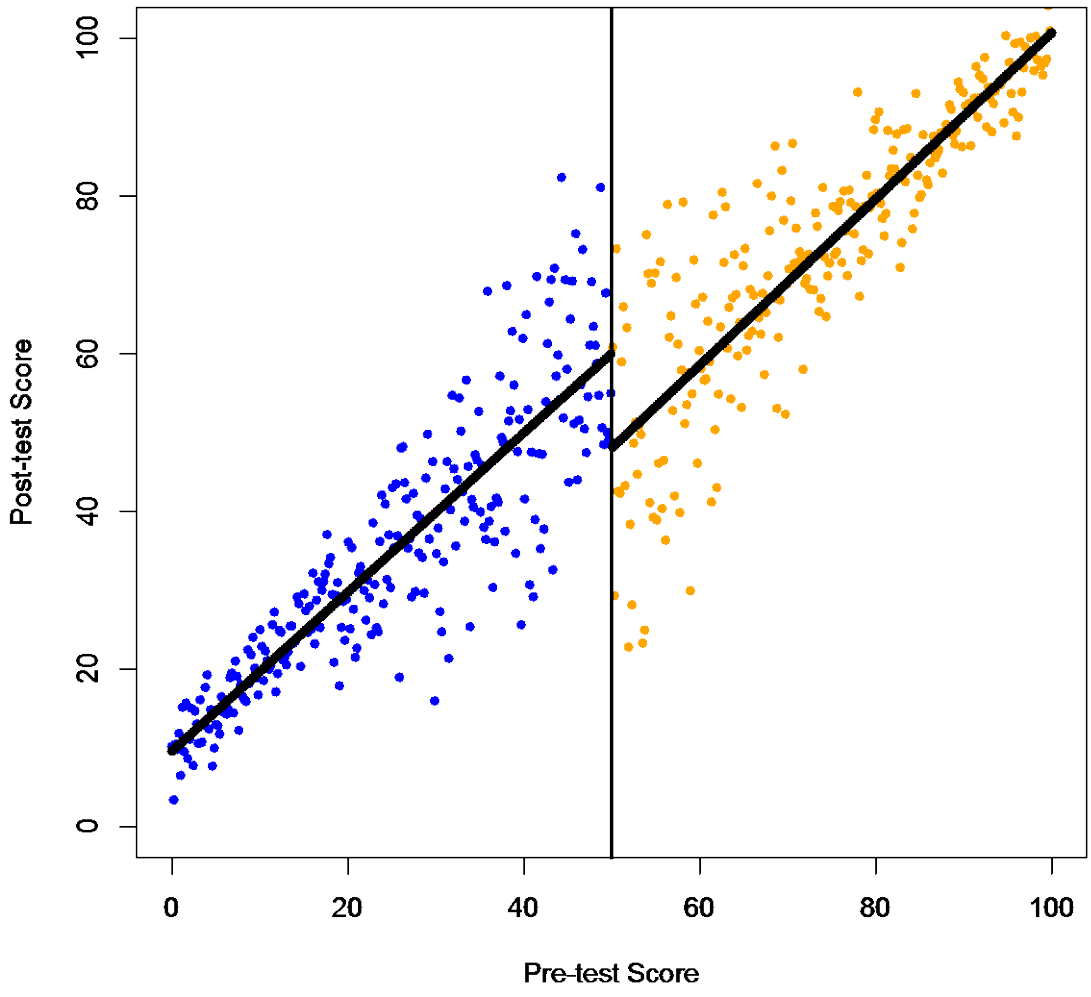
1. **High-ability students assigned to a control group in a charter school evaluation move to a private school.** In a study of charter schools that relies on administrative data from school districts for test score outcomes, some parents whose children are not accepted into the charter school through a randomized lottery might look for opportunities to move their children into a private school outside of the study. This reaction to the lottery could result in the best students leaving the control group but not the treatment group, creating the illusion of a positive impact.
2. **Teachers in the treatment group discourage low-ability students from taking an achievement test.** In a study of financial incentives for teachers whose students show the highest performance gains, teachers in the treatment group might have an incentive to discourage low-ability students from taking the test used to measure the teacher's performance.
3. **A dropout prevention program keeps lower-ability students in school in the treatment group, resulting in biased impacts on academic achievement outcomes.** By design, a dropout prevention program is intended to affect whether students remain in school, which in turn can affect attrition since dropouts often have missing data. If the program is successful, then the treatment group may include students who would have dropped out had they been in the control group. This phenomenon could result in a different mix of students taking achievement tests in the treatment and control groups.

The bottom line is there are compelling reasons for researchers to continue conducting studies that are powered to detect small impacts, but researchers should be more attuned to the threat of attrition bias in these studies. In order to adequately contain potential bias and the risk of making false inferences, researchers should be prepared to invest additional resources to keep attrition at levels below what is typical for many past studies that have been powered to detect small impacts. Researchers might also consider whether more optimistic assumptions about the attrition process are warranted in their study than what has been typical in prior studies of education interventions. More optimistic assumptions allow for attrition levels that are in the range of what past studies have experienced. Attrition models such as the one developed by the WWC are a useful tool for assessing both the level of attrition that is acceptable for a given set of assumptions about the attrition process, and how optimistic these assumptions need to be for a given level of attrition.

Is functional form misspecification bias more problematic in RDDs that are powered to detect small impacts?

Under an RDD, a cutoff on a continuous assignment variable is used to determine who is offered the opportunity to participate in a program. If the program has an impact, we would expect to see an abrupt change – a "discontinuity" – in the trend of the outcome at the cutoff. For example, because of funding constraints, a school district might only provide free after-school math tutoring to students scoring below a cutoff on a pre-test, creating the opportunity to estimate the impact of free math tutoring using an RDD. Students with scores below the cutoff would be in the treatment group; students above the cutoff would be in the comparison group. A valid estimate of the impact of tutoring could then be obtained by comparing the outcomes of students below and above the cutoff, after adjusting for students' scores on the pre-test. Figure 11 illustrates this example. In the figure, the cutoff is 50 and the impact of the tutoring program is to increase students' post-test score by 10 points. In this artificial example, the relationship or functional form between the outcome (post-test score) and the assignment variable (pre-test score) is linear.

Figure 11. Example of an RDD using simulated data



Note: In this example, the assignment variable is a pre-test score and the outcome is a post-test score. The cutoff on the assignment variable is 50. Students below the cutoff receive the program. The estimated impact of the program is the discontinuity in the black regression line that occurs at the cutoff.

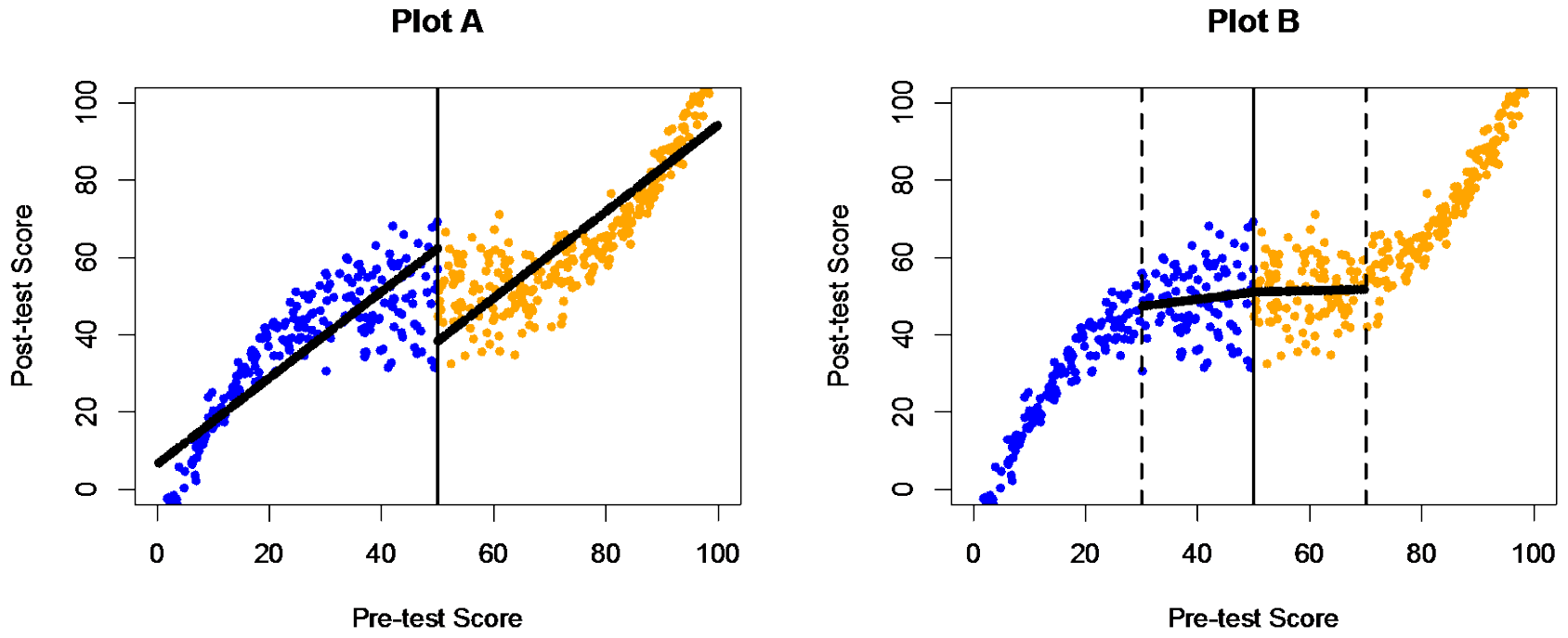
Unlike an RCT, the validity of an RDD hinges on statistical modeling, specifically modeling of the relationship between the outcome and the assignment variable. For example, if the true relationship between the outcome and the assignment variable is not linear, then fitting a linear regression line to all of the data on either side of the cutoff might result in a biased impact estimate. In Plot A of Figure 12, the discontinuity in the black regression lines is due entirely to bias resulting from the functional form misspecification (the correct functional form is cubic in this example).

State-of-the-art RDD methods address functional form misspecification bias by selecting a bandwidth (or narrow window) around the treatment–comparison cutoff and estimating a linear regression within the bandwidth (Gelman and Imbens 2014; Calonico et al. 2014; Imbens and Kalyanaraman 2012). Generally, smaller bandwidths yield less functional form misspecification bias because linear approximations become more appropriate as bandwidths get smaller. This approach is illustrated in Plot B of Figure 12, where the bandwidth is indicated by the vertical dashed lines and linear regression lines are fit using only data within the bandwidth.

However, smaller bandwidths also include fewer data points, thus adversely impacting the precision of the estimate. To manage the trade-off between bias and precision, these algorithms choose the bandwidth that minimizes the mean squared error (for example, Imbens and Kalyanaraman 2012; Calonico et al. 2014). The mean squared error is defined as the square of bias plus the variance of the impact estimate. By choosing the bandwidth that minimizes mean squared error, these algorithms select a bandwidth that yields an *asymptotically* unbiased impact estimate. As the study sample size becomes larger, the selected bandwidth becomes smaller, eventually reducing bias to zero as the sample approaches infinity. But so long as the sample size is finite, it is possible that impact estimates are biased if the true functional form within the selected bandwidth is not perfectly linear.

If, as a study becomes larger, the standard error of the impact estimate shrinks more quickly than the functional form misspecification bias (that is, precision increases much faster than bias shrinks), then Type 1 errors could become more common, even though the bias is smaller. For example, consider two studies of different sizes in which an RDD is used to test an education intervention that truly has no impact on student achievement. In one RDD study, there is a sample of 500 students, and the researcher estimates an impact of 0.06 standard deviations with a standard error of 0.04, which is not statistically significant at conventional levels. In the second RDD study, there is a larger sample of 5,000 students, and the researcher estimates an impact of 0.04 with a standard error of 0.02, which is statistically significant at conventional levels. In this example, the larger study is "better" in the sense that the bias in the impact estimate is smaller (0.04 versus 0.06 relative to a true null impact). On the other hand, the larger study is also "worse" because it leads to a Type 1 error.

Figure 12. Example of functional form misspecification bias in an RDD using simulated data



Note: In this example, the true relationship between the outcome and the assignment variable is cubic and the true impact of the program is zero. Plot A illustrates the functional form misspecification bias resulting from a linear regression using all of the data. The bias is the vertical distance between the two black regression lines. In Plot B, a linear functional form is used within a bandwidth around the cutoff. Within the bandwidth, the linear functional form is approximately correct and there is no noticeable bias.

To answer our second research question about how misspecification bias changes as the size of an RDD study increases, we use Monte Carlo simulations to assess bias under varying assumptions regarding the true relationship between the outcome and assignment variable. We examine whether statistical power increases with sample size, as well as how the magnitude of functional form misspecification bias changes and how the Type 1 error rate changes. We also assess the extent to which the technique for adjusting standard errors that Calonico et al. (2014) suggest controls Type 1 errors at the desired rate. The purpose of these exercises is to better understand the extent to which misspecification bias increases the risk of making false inferences in studies that are powered to detect small impacts, and how this potential problem might be mitigated. The purpose of these exercises is *not* to suggest that all else equal, researchers should prefer smaller studies to larger ones, as there are many cases where a larger sample size is needed to detect impacts of meaningful magnitude.

Methodological approach

Our methodological approach is to use Monte Carlo simulations, where we randomly generate data and then estimate RDD impacts, standard errors, and *p*-values using several approaches from the methodological literature. After repeating this process many times, we assess how the different approaches perform under a variety of realistic conditions that education researchers may face when conducting evaluations using an RDD.

In our Monte Carlo simulations, we generate data using seven different data generating processes (DGPs). To make the simulations findings relevant to education researchers, the DGPs are based on data from previous education studies that included math and reading post-tests and pre-tests. Each DGP consists of a fifth-order polynomial equation that describes the relationship between the assignment variable (pre-test) and the outcome (post-test). The cutoff used in each case is the median value of the pre-test. Each DGP also describes the distribution of the assignment variable, including whether and how individuals are clustered within unique values of the assignment variable. Finally, each DGP specifies what proportion of the variance of the outcome is due to the assignment variable versus unobserved random factors. Details regarding the DGPs are reported in the appendix.

The specific steps of our simulation procedure are as follows:

1. Generate data using one of the DGPs specified in the appendix. We generate data with 1,000, 10,000, or 100,000 observations.
2. Estimate RDD impacts on the simulated data using a number of state-of-the-art RDD methods. These methods include two different bandwidth selection algorithms and two different approaches to calculating standard errors. The two bandwidth algorithms are those suggested by Imbens and Kalyanaraman (2012) and Calonico et al. (2014).[10] The standard error estimation approaches are (1) a "conventional" approach that ignores finite sample bias and (2) an approach that uses Calonico et al.'s method for calculating bias-corrected impact estimates and robust standard errors.
3. Repeat steps 1 and 2 10,000 times, recording impacts, standard errors, *p*-values, and bandwidth estimates.

With thousands of simulated impact estimates, we can look at summary statistics of how the estimates perform under varying conditions. We report three sets of findings for all simulations:

1. **The mean MDE size across Monte Carlo replications assuming 80 percent power.** Using the standard error estimate for each replication, we calculate the smallest impact that would, with high

[10] In some cases, the algorithms select bandwidths that are so narrow there are not enough data to calculate an impact and/or a standard error. In those cases, we automatically expand the bandwidth until we can calculate an impact and standard error.

probability, be statistically significant at the 5 percent level given that standard error—this is just 2.8 times the standard error estimate.

2. **The mean functional form misspecification bias across Monte Carlo replications.** Because data are generated under the null hypothesis of no "true" impact, the mean bias is equal to the mean estimated impact.

3. **The Type 1 error rate across Monte Carlo replications.** This is the proportion of statistically significant impact estimates (that is, where the *p*-value of the impact is less than 0.05). In an empirical approach with appropriate inference, this false inference rate should be 0.05, as the model assumes no "true" impact.

Simulation findings

Our simulation findings show that with conventional estimation Type 1 error rates go up as studies are powered to detect smaller impacts, but that the robust estimation approach that Calonico et al. (2014) recommend solves this problem. In Table 7, we report a summary of findings in which we average across the seven DGPs (complete findings for each individual DGP are reported in the appendix). We report the expected bias, MDE, and Type 1 error rate for studies with 1,000, 10,000, or 100,000 observations. We report these statistics for both conventional and robust estimation, and for bandwidths selected using either Imbens and Kalyanaraman's (2012) or Calonico et al.'s (2014) algorithms. Note that Calonico et al.'s bandwidth selection algorithm is distinct from the robust estimation procedures that the authors also recommend.

Conventional estimation findings. As sample size increases, the MDE gets smaller, as expected. Bias also gets smaller, however at a slower rate than the MDE, resulting in an increasing rate of Type 1 errors. With Calonico et al.'s (2014) bandwidth selection algorithm and a sample size of 1,000, bias is 0.01 standard deviations, the MDE is 0.543 standard deviations, and the Type 1 error rate is 0.059. With a sample size of 100,000, bias falls to 0.006 standard deviations, the MDE falls to 0.061 standard deviations, and the Type 1 error rate increases to 0.105. The pattern of findings using Imbens and Kalyanaraman's (2012) bandwidth selection algorithm is similar.

Robust estimation findings. As sample size increases, both bias and the MDE get smaller. In this case, the rate of decline is roughly equal, and the Type 1 error rate is not adversely affected. With Calonico et al.'s (2014) bandwidth selection algorithm and a sample size of 1,000, bias is 0.004 standard deviations, the MDE is 0.683 standard deviations, and the Type 1 error rate is 0.052. With a sample size of 100,000, bias falls to 0.002 standard deviations, the MDE falls to 0.081 standard deviations, and the Type 1 error rate is 0.053. The pattern of findings using Imbens and Kalyanaraman's (2012) bandwidth selection algorithm is similar.

Table 7. Summary of findings from simulations based on data from education studies

Standard Errors	Bandwidth Selection Algorithm and Sample Size					
	Imbens and Kalyanaraman (2012)			Calonico et al. (2014)		
	1,000	10,000	100,000	1,000	10,000	100,000
Average absolute bias (in standard deviations)						
Conventional	0.012	0.010	0.007	0.010	0.007	0.006
Robust	0.003	0.003	0.003	0.004	0.003	0.002
Average minimum detectable effect (in standard deviations)						
Conventional	0.469	0.153	0.053	0.543	0.173	0.061
Robust	0.866[a]	0.256[a]	0.076	0.683	0.220	0.081

Standard Errors	Bandwidth Selection Algorithm and Sample Size					
	Imbens and Kalyanaraman (2012)			Calonico et al. (2014)		
	1,000	10,000	100,000	1,000	10,000	100,000
	Average Type 1 error rate (the target is 0.05)					
Conventional	0.060	0.067	0.114	0.059	0.059	0.105
Robust	0.051	0.049	0.052	0.052	0.050	0.053

Source: Monte Carlo simulations, 10,000 replications.

Note: The findings reported in this table are averaged across seven Monte Carlo simulations corresponding to seven data-generating processes (DGPs). The robust estimation approach included bias-corrected point estimates and standard errors inflated to control the coverage error rate as suggested by Calonico et al. (2014).

[a] Three out of the 10,000 Monte Carlo replications for one of the seven DGPs yielded extremely large standard errors that severely skewed these values. Those extreme outliers were removed from the calculation of this average minimum detectable effect.

Discussion

In this paper, we have demonstrated that although it is often necessary to conduct a study capable of detecting small impacts, researchers should be aware of the greater risk of false inference due to small biases. This concern applies to two of the strongest possible evaluation designs—the RCT and RDD. We have shown that as studies are powered to detect smaller impacts, some types of bias that previously might have been negligible can become significant threats to the credibility of a study's findings. This is because although statistical power generally increases with sample size, some sources of bias *do not decrease* with sample size (in the case of attrition bias in RCTs) or do not always decrease as quickly as power increases (in the case of functional form misspecification bias in RDDs). Thus, the *relative* threat of these biases can become larger in studies that are powered to detect smaller impacts. Fortunately, with proper awareness and action, researchers can likely mitigate these threats. Below we discuss strategies that researchers can consider using.

Interpretive Caution

Readers should *not* interpret this report's findings as supporting the design of underpowered studies. All else being equal, more statistical power is always better. For example, in a study that seeks to detect an impact of 0.10 standard deviations, it would be better to have 80% power than to have 60% power.

Strategies to address small biases due to attrition in RCTs

In the case of attrition bias in RCTs, we suggest three strategies. First, researchers can mitigate bias by expending more resources to achieve higher response rates for the collection of outcome data. However, even with substantially greater study resources, it might not always be possible to reduce attrition to the extent necessary since there may be diminishing marginal returns for each additional dollar invested in reducing attrition.

Second, attrition bias could be partially mitigated in some studies by statistically adjusting for observed differences in baseline characteristics between those who do and do not attrit, and how that difference varies between the treatment and control groups. Puma et al (2009) examine several different approaches to accounting for missing outcome data including multiple imputations, regression adjustment, and nonresponse weights. These analytic adjustments will be most effective when researchers have access to baseline data that are both correlated with outcomes and correlated with attrition.[11]

[11] The WWC attrition model does not directly incorporate covariates. However, the benefits of adjusting for covariates can be reflected in the model by making more optimistic assumptions regarding the negative consequences of attrition.

Third, in some contexts researchers might be able to make more optimistic assumptions regarding the negative consequences of attrition. Attrition models, such as the one developed by the WWC and used in federal evidence reviews, can provide a framework for incorporating these assumptions into an assessment of acceptable levels of attrition. However, more optimistic assumptions should only be made when appropriate for the study context. More optimistic assumptions that are unwarranted may lead to a low-quality study with misleading findings.

When considering the second strategy, researchers could conduct an empirical check to examine how correlated attrition is with baseline measures of the outcome variables, which serve as a proxy for actual outcomes. Researchers could also examine the differences in baseline characteristics between attriters and non-attriters, as well as differences in baseline characteristics between attriters from the treatment group and attriters from the control group. However, we strongly caution against taking these empirical checks as absolute truth, as they may not be precisely estimated in many studies. We therefore also recommend supplementing any empirical checks with an intentional theory for why a particular intervention may or may not have a strong influence on attrition. For example, it is arguably less plausible that an intervention focused on increasing physical activity during recess would have a strong impact on attrition; on the other hand, it might be more plausible that whether a student is admitted to a charter school has a noticeable effect on whether the student chooses to enroll in a private school and hence has missing outcome data.

Strategies to address small biases due to functional form misspecification in RDDs

In the case of functional form misspecification bias in RDDs, we can avoid mistakes in inference if we use existing methods to inflate standard errors (Calonico et al. 2014). However, this correction does increase sample size requirements. In some cases, the inflation of standard errors can make it practically impossible to detect impacts smaller than 0.05 standard deviations.

We also suggest that researchers consider whether an RDD study is the most appropriate method for a particular education intervention. If the relationship between the assignment variable and outcomes is likely to be highly non-linear or if the assignment variable is very lumpy (see Appendix Figures A1-A7 for examples), then it might be impossible to detect meaningful small impacts using an RDD, even if large sample sizes are available. In these cases, researchers should consider alternate methods for evaluating the intervention, such as RCTs.

Strategies to address small biases in all study designs

We conclude by offering a few general suggestions for researchers to consider as they plan and implement future impact studies. Our first suggestion is to reemphasize a point made by other researchers: during the planning stages of a study, researchers should be thoughtful about what is a reasonable target minimum detectable effect for the particular intervention tested. At minimum researchers should consider how much the intervention costs. (Other factors that might be relevant are the impacts similar interventions have obtained in the past, how impacts would compare to existing policy-relevant performance gaps, and how impacts would compare to typical academic growth trajectories). For instance, smaller impacts could still be substantively important if the cost of the intervention for the average student is relatively small; in that case a larger sample might be appropriate in order to achieve a small minimum detectable effect. By contrast, an intervention that requires a large investment for each student served may not require a small minimum detectable effect, since the cost of implementing the intervention would only be justified if it was found to have a very large effect. Given the costs of conducting studies that are powered to detect small impacts and the increased risks of false inference due to small biases, researchers should be able to articulate an intentional argument as to why a small impact is important to detect in each particular context.

A second, related suggestion is that, regardless of the sample sizes selected, researchers should have a compelling theory of action relating proximal outcomes to distal outcomes, and they should ideally

collect data on both of these outcomes. This is because a small impact on a distal outcome may be more credible if it is accompanied by a large impact on a logically connected proximal outcome. Typically, impacts are larger on proximal outcomes, and in some cases, proximal outcomes might be of intrinsic interest. For example, a text-messaging program to students might have a proximal goal of increasing attendance and a distal goal of increasing college enrollment. Because attendance itself is a behavioral outcome of interest to many schools, focusing on this outcome could allow for a more modestly powered study without sacrificing policy relevance. That said, we recognize in many cases, there is policy interest in distal outcomes such as student achievement and high school graduation. In these cases, in which studies need high statistical power to detect small distal impacts, we suggest that researchers still collect information on proximal outcomes to accompany the distal outcomes. We encourage researchers to show a strong theoretical and empirical link between the proximal and distal outcomes to help protect against potentially spurious impacts. At a minimum, if researchers find a statistically significant small impact on the distal outcome, they should be able to show that there are also larger impacts on the proximal outcome and that the proximal and distal outcomes are strongly correlated.

Ultimately, we cannot offer any single solution for addressing these challenges—the best approach is likely to vary by context. However, we do recommend that researchers resist the temptation to ignore these "small" biases. Even if these biases cannot be fully addressed, they can at least be acknowledged and mitigated to the extent possible. Consumers of research can then make more informed decisions about how much weight to put on the impact findings when making high-stakes decisions.

Appendix

This appendix provides additional details regarding the data generating processes (DGPs) used in the RDD Monte Carlo simulations. It also provides more detailed findings for each individual DGP (the findings in the main text are aggregated across all DGPs).

In our Monte Carlo simulations, we generate data using seven different DGPs. To make the simulations findings relevant to education researchers, the DGPs are based on data from previous education studies that included math and reading post-tests and pre-tests. The data sources are described in Table A1.

Each DGP consists of a fifth-order polynomial equation that describes the relationship between the assignment variable (pre-test) and the outcome (post-test). The coefficients in the models were estimated using the data sources described in Table A1. We report the coefficient estimates in Table A2.

Each DGP also describes the distribution of the assignment variable, including whether and how individuals are clustered within unique values of the assignment variable. These distributions were empirically estimated using the data sources described in Table A1. We report the empirical distributions in Tables A3-A9.

Visualizations of these data generating processes are shown in Figures A1-A7. In each figure, randomly generated data points are plotted along with the polynomials described in Table 3. The frequencies of the data points follow the empirical distributions reported in Tables A3-A9.

We report full simulation results for each model in Tables A10-A12. In Table A10 we report the bias for each model, in Table A11 we report the minimum detectable effects, and in Table A12 we report Type 1 error rates.

Table A1. Data from past evaluations used in simulations

Study	Purpose	Student Grade	Student Outcome Measures	Unit of Random Assignment	Number of States	Number of Districts	Number of Schools	Number of Students
Evaluation of Reading Comprehension Interventions (James-Burdumy et al. 2010)	This study evaluates the impact of four interventions on fifth-grade reading achievement.	5	Group Reading Assessment and Diagnostic Evaluation (GRADE)	School	8	10	90	6,350
Evaluation of Teacher Preparation Models (Constantine et al. 2009)	This study examines the impact of different approaches to teacher preparation on teacher practice and student performance.	K-5	Reading Comprehension, Vocabulary, and Math Concepts and Applications subtests of the California Achievement Tests, 5th Edition	Student	7	20	60	2,490
Evaluation of the Effectiveness of Reading and Mathematics Software Products (EERMSP) (Campuzano et al. 2009)	This study randomly assigned teachers to a treatment group that uses a specified educational technology, or a control group that used conventional teaching approaches. The study consisted of four sub-studies of different interventions at different grade levels (see three rows below).	-	-	-	-	-	-	-
EERMSP Grade 1	-	1	Stanford Achievement Test (version 10) Reading, and Test of Word Reading Efficiency	Teacher	12	20	50	4,420
EERMSP Grade 4	-	4	Stanford Achievement Test (version 10), Reading	Teacher	9	10	40	3,110
EERMSP Grade 6	-	6	Stanford Achievement Test (version 10), Math	Teacher	7	10	30	4,260

Source: Randomized controlled trials previously completed by Mathematica for IES.

Note: Student, district, and school sample sizes are rounded to the nearest 10 in accordance with NCES publication policy. State sample sizes are taken from the citations listed in the first column.

Table A2. Polynomial regression results

Study	Test Score	Number of Unique Pre-test Values	Regression Coefficients						Adj- R^2
			b0	b1	b2	b3	b4	b5	
Evaluation of the Effectiveness of Reading and Mathematics Software Products (Campuzano et al. 2009)	SAT-10 Reading (grade 1)	95	27.7	2.16	-0.111	0.00229	$-2.02*10^{-5}$	$6.57*10^{-8}$	0.21
	SAT-10 Reading (grade 4)	58	47.3	-3.65	0.202	-0.00516	$6.66*10^{-5}$	$-3.31*10^{-7}$	0.27
	SAT-10 Math (grade 6)	46	56.1	-6.20	0.400	-0.133	$2.37*10^{-4}$	$-1.71*10^{-6}$	0.16
Evaluation of Teacher Preparation Models (Constantine et al. 2009)	CAT-5 Vocabulary	96	25.3	0.562	-0.031	0.00109	$-1.31*10^{-5}$	$5.33*10^{-8}$	0.18
	CAT-5 Math	99	27.5	-0.124	0.0177	-0.000177	$9.90*10^{-7}$	$-3.22*10^{-9}$	0.21
	CAT-5 Reading Comprehension	91	21.5	1.87	-0.110	0.00291	$-3.10*10^{-5}$	$1.15*10^{-7}$	0.17
Evaluation of Reading Comprehension Interventions (James-Burdumy et al. 2010)	GRADE	31	2510	-137	2.96	-0.0311	$1.59*10^{-4}$	$-3.17*10^{-7}$	0.19

Source: Data from the restricted-use files corresponding to the listed studies.

Note: This table reports coefficients from a regression of scores on the specified test at follow-up on a fifth-order polynomial of scores from the same test administered at baseline. Specifically, the regression is $y = b0 + b1 \bullet x + b2 \bullet x^2 + b3 \bullet x^3 + b4 \bullet x^4 + b5 \bullet x^5$, where y is the follow-up test score and x is the baseline test score.

CAT-5 = California Achievement Tests, 5th Edition

GRADE = Group Reading Assessment and Diagnostic Evaluation

SAT-10 = Stanford Achievement Test (version 10)

Table A3. Relative frequency by unique value of the assignment variable, SAT-10 Reading (grade 1)

Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency
0	0.0002	36	0.0019	56	0.0072	76	0.0132	96	0.018
11	0.0002	37	0.0026	57	0.0103	77	0.0123	97	0.0178
15	0.0002	38	0.0024	58	0.0075	78	0.0123	98	0.019
18	0.0005	39	0.0038	59	0.0106	79	0.0144	99	0.0226
19	0.0005	40	0.0046	60	0.0099	80	0.0142	100	0.019
21	0.0002	41	0.0034	61	0.0091	81	0.0163	101	0.0224
22	0.0002	42	0.0043	62	0.0123	82	0.0123	102	0.0267
23	0.0002	43	0.0031	63	0.0118	83	0.0135	103	0.0228
24	0.0002	44	0.0041	64	0.0118	84	0.0151	104	0.0274
25	0.0005	45	0.005	65	0.0123	85	0.0113	105	0.0245
26	0.0002	46	0.005	66	0.0118	86	0.0147	106	0.0281
27	0.001	47	0.0072	67	0.0115	87	0.0175	107	0.03
28	0.001	48	0.0065	68	0.0118	88	0.0123	108	0.0356
29	0.001	49	0.0058	69	0.0142	89	0.012	109	0.0255
30	0.0002	50	0.006	70	0.0142	90	0.0171	110	0.0207
31	0.0024	51	0.0067	71	0.0084	91	0.0192		
32	0.0017	52	0.0072	72	0.0149	92	0.0175		
33	0.0012	53	0.0079	73	0.0137	93	0.012		
34	0.0019	54	0.0089	74	0.0132	94	0.0159		
35	0.0026	55	0.0103	75	0.0135	95	0.0137		

Source: Data from the restricted-use file for Evaluation of the Effectiveness of Reading and Mathematics Software Products (Campuzano et al. 2009).

Note: This table reports the relative frequency for each unique value of the variable.

SAT-10 = Stanford Achievement Test (version 10)

Table A4. Relative frequency by unique value of the assignment variable, SAT-10 Reading (grade 4)

Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency
12	0.0004	25	0.0151	37	0.0241	49	0.027	61	0.0187
14	0.0011	26	0.0108	38	0.0187	50	0.0241	62	0.0173
15	0.0018	27	0.0176	39	0.027	51	0.0259	63	0.0176
16	0.0011	28	0.0151	40	0.0262	52	0.0259	64	0.0144
17	0.0032	29	0.0187	41	0.0259	53	0.0316	65	0.0165
18	0.0018	30	0.0173	42	0.0298	54	0.0208	66	0.0075
19	0.0029	31	0.0208	43	0.0259	55	0.0248	67	0.0075
20	0.0104	32	0.0194	44	0.0288	56	0.0252	68	0.0036
21	0.0093	33	0.0191	45	0.0234	57	0.0262	69	0.0022
22	0.0086	34	0.0208	46	0.0288	58	0.0194	70	0.0014
23	0.0147	35	0.0176	47	0.0298	59	0.0219		
24	0.0097	36	0.0234	48	0.0262	60	0.0252		

Source: Data from the restricted-use file for Evaluation of the Effectiveness of Reading and Mathematics Software Products (Campuzano et al. 2009).

Note: This table reports the relative frequency for each unique value of the variable.

SAT-10 = Stanford Achievement Test (version 10)

Table A5. Relative frequency by unique value of the assignment variable, SAT-10 Reading (grade 6)

Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency
5	0.001	15	0.0195	25	0.0281	35	0.0308	45	0.018
6	0.0012	16	0.0259	26	0.0316	36	0.0281	46	0.0141
7	0.0025	17	0.0232	27	0.0306	37	0.0318	47	0.0126
8	0.0032	18	0.0294	28	0.0262	38	0.0276	48	0.0054
9	0.0052	19	0.0289	29	0.0269	39	0.0303	49	0.0074
10	0.0113	20	0.0313	30	0.0331	40	0.0318	50	0.0025
11	0.0099	21	0.0338	31	0.0308	41	0.0269		
12	0.0126	22	0.0276	32	0.0313	42	0.0274		
13	0.017	23	0.0311	33	0.0244	43	0.0222		
14	0.022	24	0.0299	34	0.0318	44	0.0217		

Source: Data from the restricted-use file for Evaluation of the Effectiveness of Reading and Mathematics Software Products (Campuzano et al. 2009).

Note: This table reports the relative frequency for each unique value of the variable.

SAT-10 = Stanford Achievement Test (version 10)

Table A6. Relative frequency by unique value of the assignment variable, CAT-5 Vocabulary

Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency
1	0.0383	21	0.0102	41	0.0214	61	0.0137	81	0.0028
2	0.0014	22	0.0126	42	0.0168	62	0.0123	82	0.0025
3	0.0049	23	0.0074	43	0.0088	63	0.0102	83	0.0028
4	0.0056	24	0.013	44	0.0197	64	0.0112	84	0.0028
5	0.0049	25	0.0147	45	0.0172	65	0.0119	85	0.0028
6	0.0046	26	0.0095	46	0.0154	66	0.0119	86	0.0025
7	0.0084	27	0.0165	47	0.0207	67	0.0105	87	0.0025
8	0.0053	28	0.0144	48	0.0221	68	0.0042	88	0.0004
9	0.0088	29	0.0112	49	0.0154	69	0.0081	89	0.0007
10	0.0074	30	0.0144	50	0.0133	70	0.0081	90	0.0014
11	0.0144	31	0.0176	51	0.0256	71	0.0084	91	0.0018
12	0.007	32	0.0095	52	0.0126	72	0.007	94	0.0007
13	0.0098	33	0.019	53	0.0235	73	0.0049	95	0.0014
14	0.0077	34	0.0193	54	0.0232	74	0.0035	96	0.0004
15	0.0102	35	0.0137	55	0.0098	75	0.0025	98	0.0014
16	0.007	36	0.0193	56	0.0158	76	0.0032	99	0.0046
17	0.0095	37	0.0207	57	0.0186	77	0.0056		
18	0.0133	38	0.0151	58	0.0154	78	0.0035		
19	0.0095	39	0.0168	59	0.0102	79	0.0025		
20	0.0144	40	0.0176	60	0.0109	80	0.0042		

Source: Data from the restricted-use file for Evaluation of Teacher Preparation Models (Constantine et al. 2009)

Note: This table reports the relative frequency for each unique value of the variable.

CAT-5 = California Achievement Tests, 5th Edition

Table A7. Relative frequency by unique value of the assignment variable, CAT-5 Math

Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency
1	0.0478	21	0.0122	41	0.0194	61	0.014	81	0.0032
2	0.0043	22	0.0097	42	0.0198	62	0.0086	82	0.004
3	0.0032	23	0.0086	43	0.0216	63	0.0126	83	0.0029
4	0.004	24	0.0101	44	0.0237	64	0.009	84	0.0011
5	0.0029	25	0.0072	45	0.0198	65	0.0079	85	0.0007
6	0.0029	26	0.0104	46	0.0205	66	0.009	86	0.0018
7	0.004	27	0.0079	47	0.0176	67	0.009	87	0.0004
8	0.0065	28	0.0065	48	0.0205	68	0.0119	88	0.0036
9	0.009	29	0.018	49	0.0194	69	0.0101	89	0.0011
10	0.0072	30	0.0147	50	0.018	70	0.0093	90	0.0029
11	0.0065	31	0.0119	51	0.0194	71	0.0075	91	0.0004
12	0.0075	32	0.0162	52	0.0183	72	0.0093	92	0.0011
13	0.0057	33	0.0147	53	0.0198	73	0.0086	93	0.0029
14	0.0068	34	0.018	54	0.0216	74	0.0036	94	0.0014
15	0.0083	35	0.0162	55	0.0133	75	0.0079	95	0.0007
16	0.0054	36	0.0136	56	0.014	76	0.0065	96	0.0014
17	0.0061	37	0.0194	57	0.0136	77	0.0057	97	0.0004
18	0.0083	38	0.0169	58	0.018	78	0.0057	98	0.0004
19	0.0097	39	0.0133	59	0.0165	79	0.0047	99	0.0086
20	0.0083	40	0.0223	60	0.0122	80	0.0018		

Source: Data from the restricted-use file for Evaluation of Teacher Preparation Models (Constantine et al. 2009)

Note: This table reports the relative frequency for each unique value of the variable.

CAT-5 = California Achievement Tests, 5th Edition

Table A8. Relative frequency by unique value of the assignment variable, CAT-5 Reading comprehension

Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency
1	0.0686	21	0.0103	41	0.0206	61	0.0124	81	0.0028
2	0.0067	22	0.0117	42	0.0178	62	0.0142	82	0.0021
3	0.0032	23	0.0082	43	0.0238	63	0.0089	83	0.0011
4	0.0028	24	0.0092	44	0.0181	64	0.0064	84	0.0004
5	0.0057	25	0.0142	45	0.0227	65	0.0067	85	0.0004
6	0.0075	26	0.011	46	0.0185	66	0.0082	87	0.0007
7	0.0067	27	0.0114	47	0.022	67	0.006	89	0.0007
8	0.0057	28	0.0149	48	0.0202	68	0.0064	94	0.0007
9	0.006	29	0.0131	49	0.0263	69	0.006	96	0.0007
10	0.0046	30	0.0117	50	0.0171	70	0.0057	97	0.0004
11	0.0078	31	0.0117	51	0.0227	71	0.0053	99	0.0124
12	0.0075	32	0.0146	52	0.0139	72	0.0039		
13	0.0067	33	0.0153	53	0.0231	73	0.0025		
14	0.0135	34	0.0167	54	0.0163	74	0.0025		
15	0.0075	35	0.0174	55	0.0231	75	0.0032		
16	0.0114	36	0.0245	56	0.0146	76	0.0036		
17	0.0067	37	0.0149	57	0.0149	77	0.0014		
18	0.0092	38	0.0178	58	0.0139	78	0.0025		
19	0.0078	39	0.0245	59	0.0096	79	0.0014		
20	0.0075	40	0.0188	60	0.0146	80	0.0014		

Source: Data from the restricted-use file for Evaluation of Teacher Preparation Models (Constantine et al. 2009)

Note: This table reports the relative frequency for each unique value of the variable.

CAT-5 = California Achievement Tests, 5th Edition

Table A9. Relative frequency by unique value of the assignment variable, GRADE

Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency	Value	Relative Frequency
65	0.0007	81	0.0301	96	0.0583	111	0.0406	127	0.0117
68	0.0003	83	0.0373	98	0.0586	114	0.0406	129	0.0082
70	0.0012	85	0.044	100	0.0593	116	0.0344	131	0.0026
72	0.0043	87	0.0505	103	0.0559	118	0.0325		
74	0.0065	90	0.0534	105	0.0548	120	0.0301		
76	0.0146	92	0.0497	107	0.0524	122	0.0201		
79	0.0225	94	0.0577	109	0.0473	125	0.0199		

Source: Data from the restricted-use file for Evaluation of Reading Comprehension Interventions (James-Burdumy et al. 2010)

Note: This table reports the relative frequency for each unique value of the variable.

GRADE = Group Reading Assessment and Diagnostic Evaluation

Figure A1. Visualization of data generating model for SAT-10 Reading (grade 1)

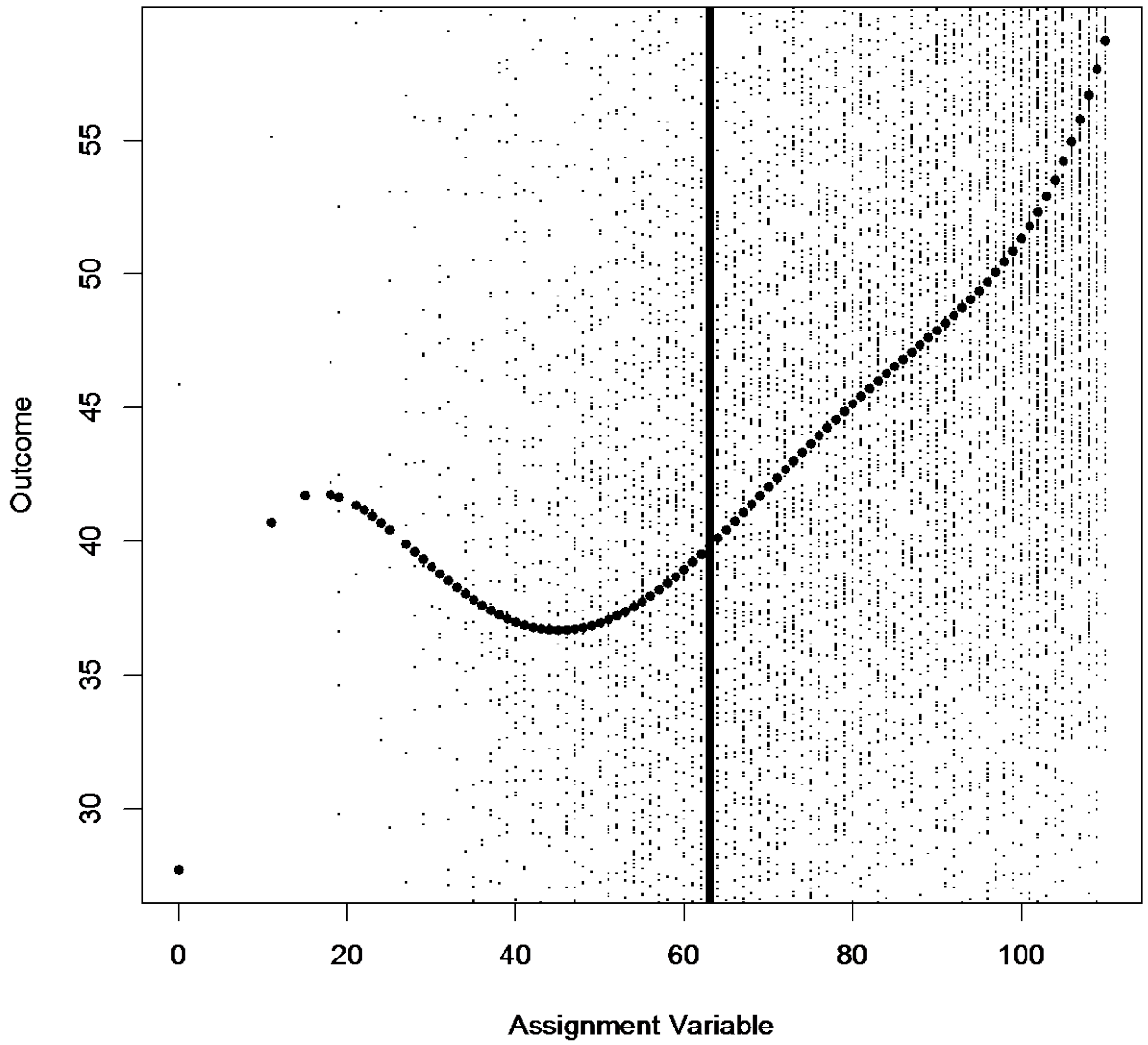


Figure A2. Visualization of data generating model for SAT-10 Reading (grade 4)

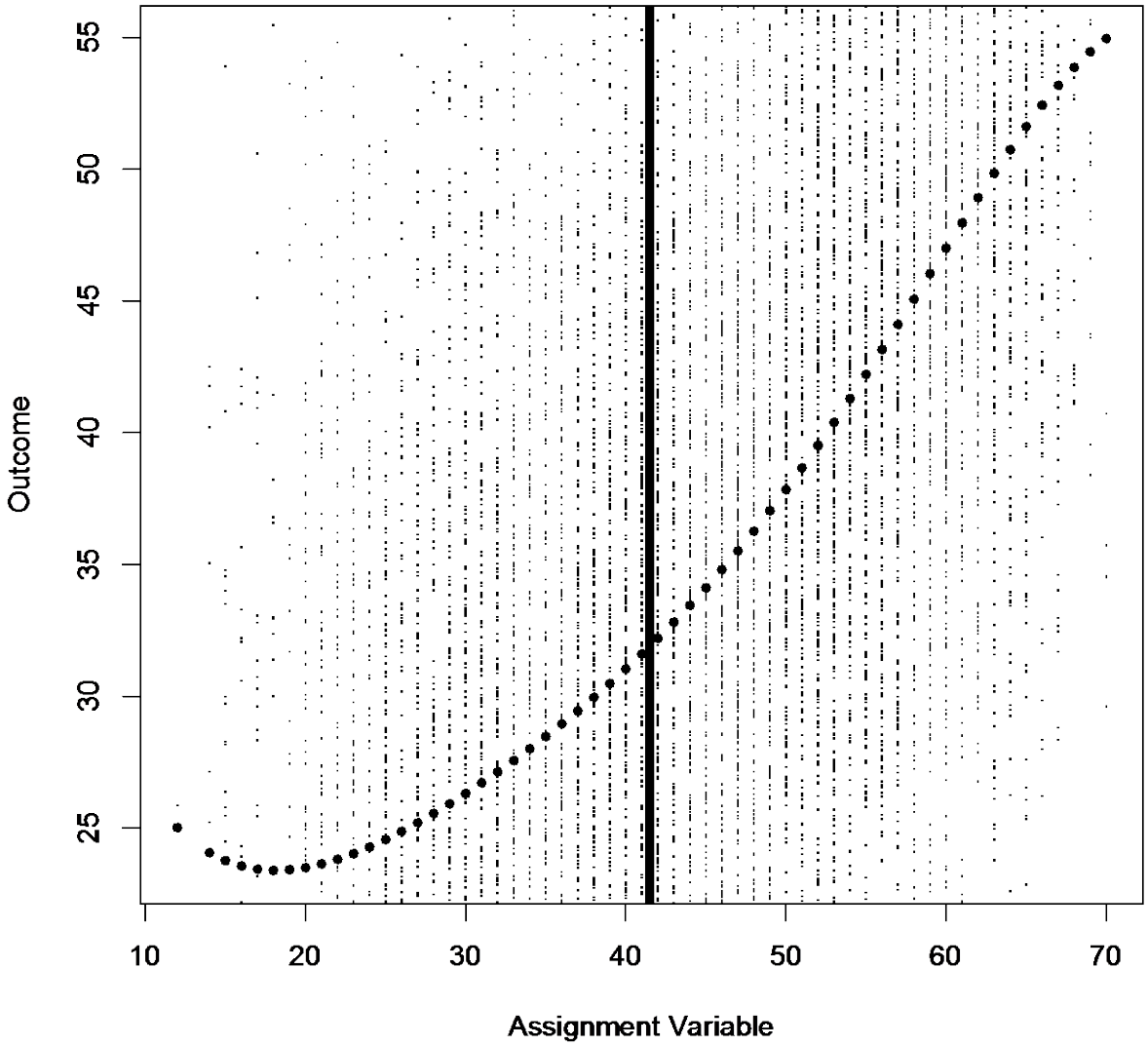


Figure A3. Visualization of data generating model for SAT-10 Math (grade 6)

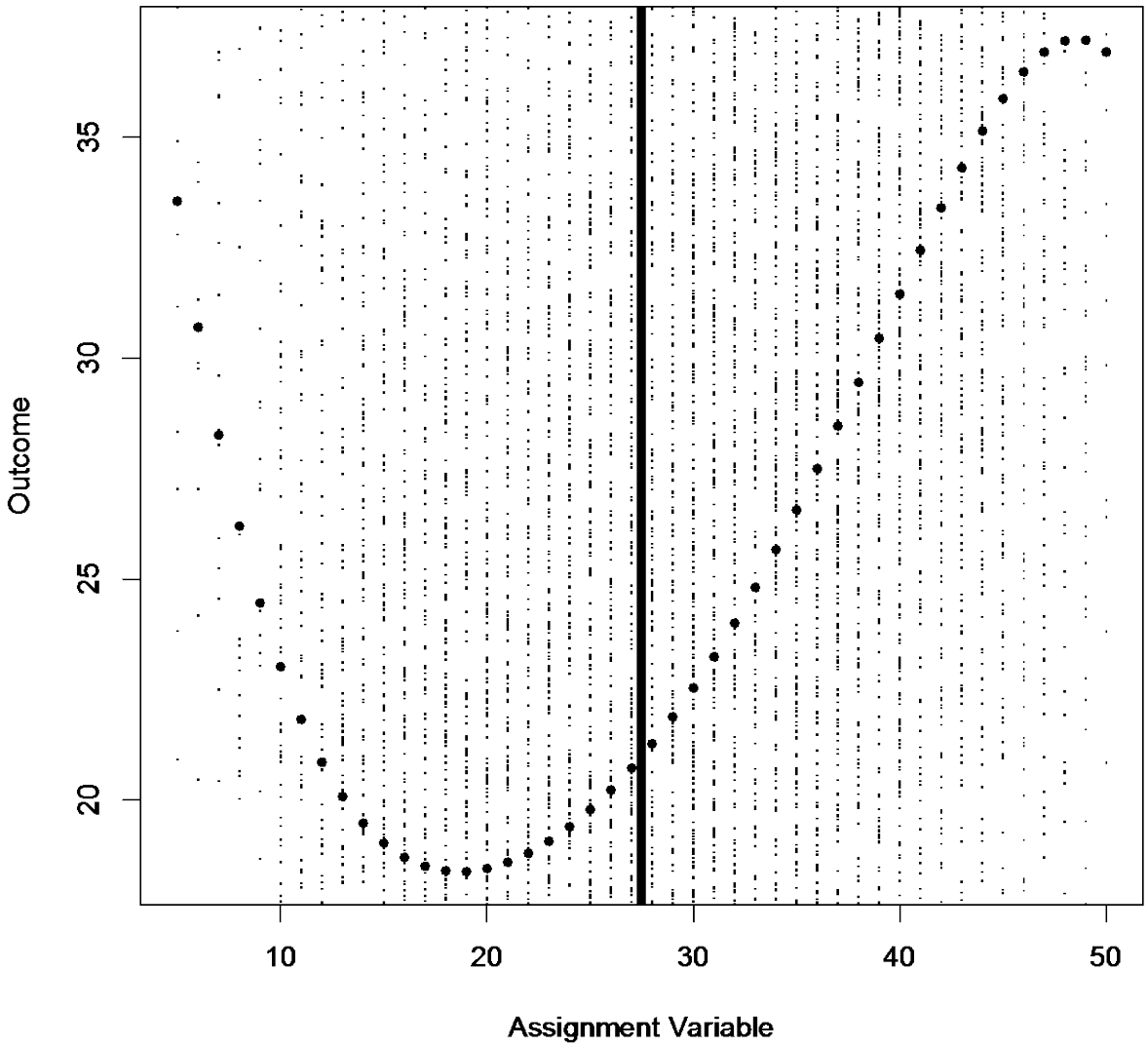


Figure A4. Visualization of data generating model for GRADE

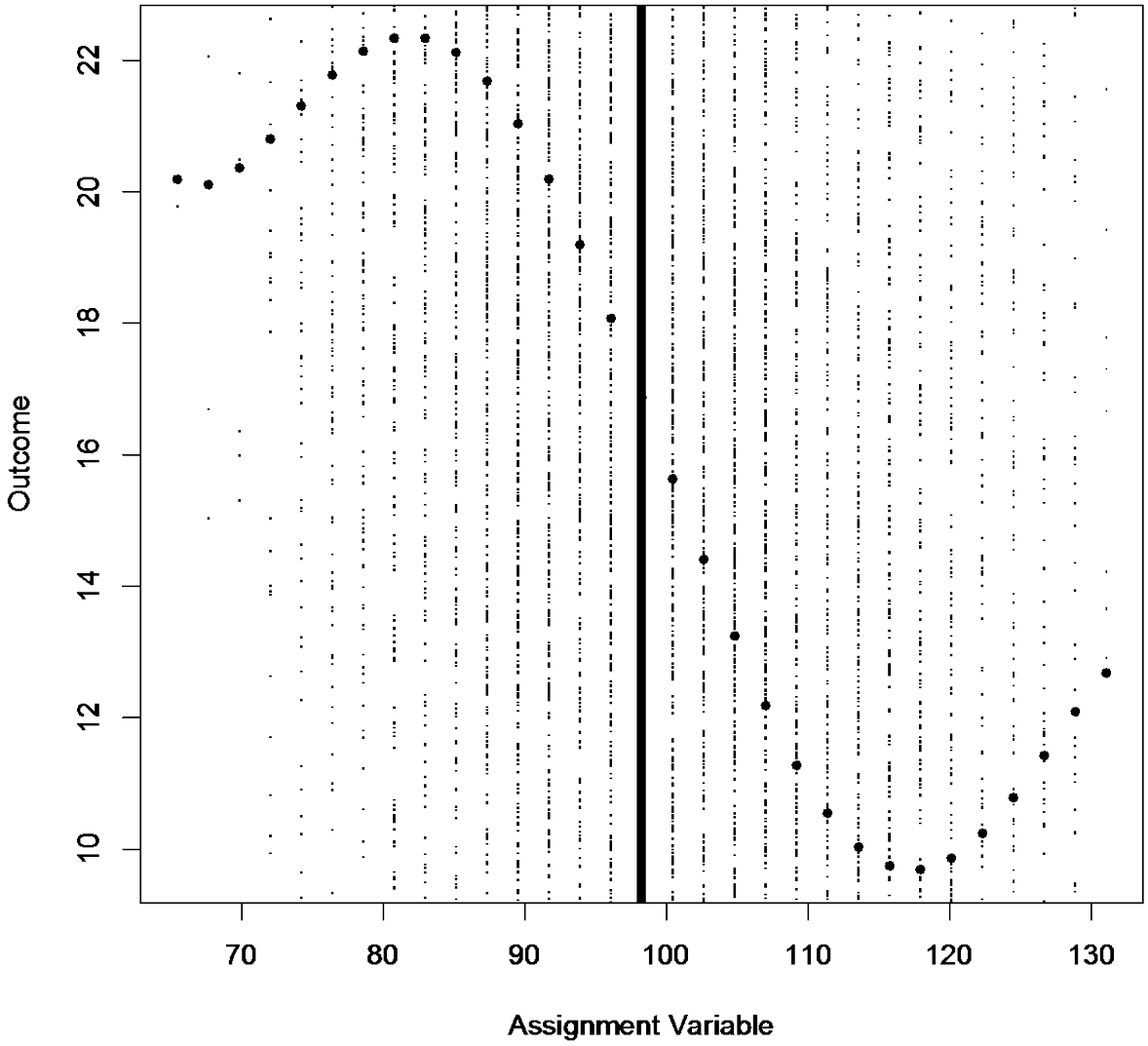


Figure A5. Visualization of data generating model for CAT-5 Vocabulary

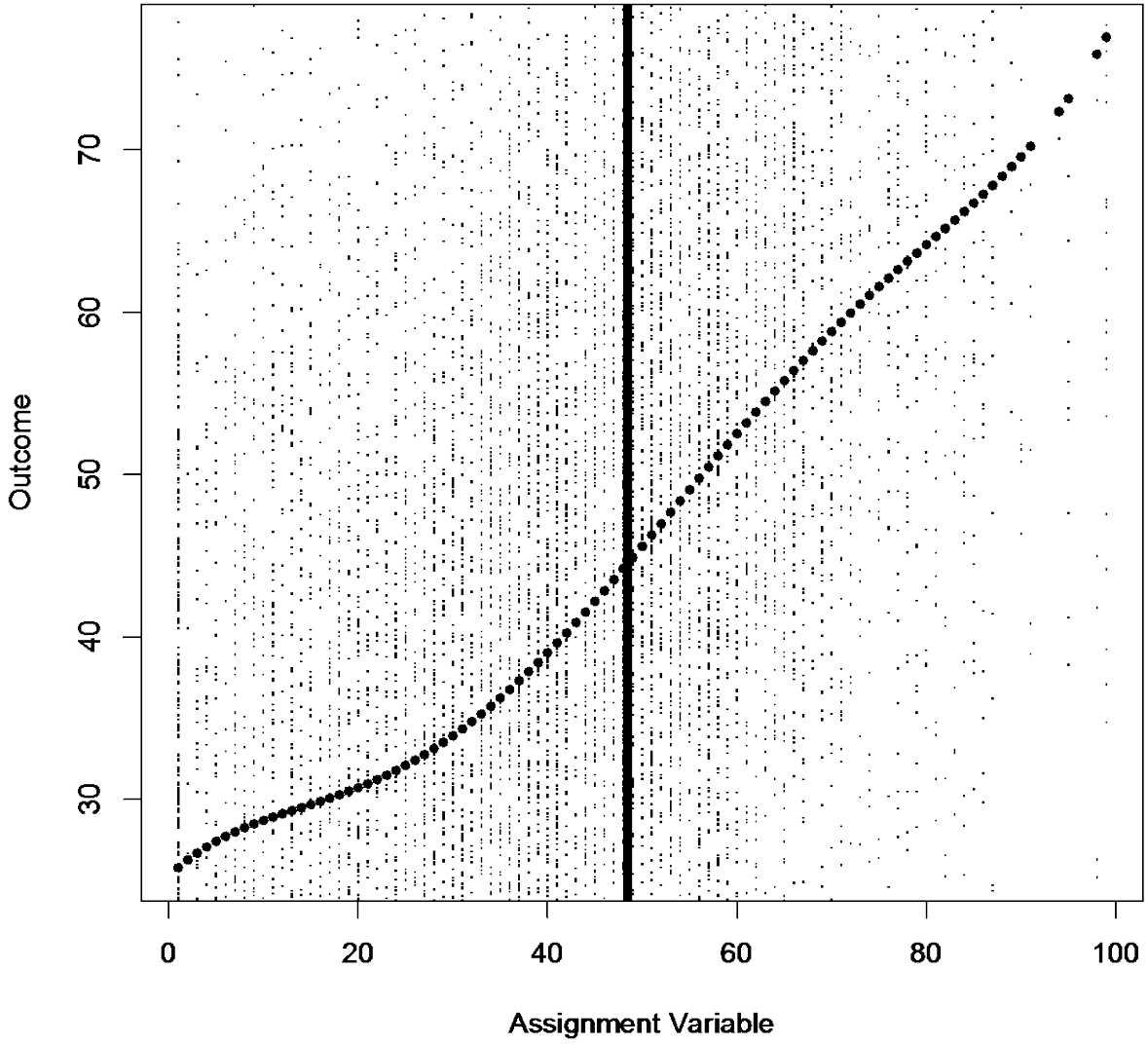


Figure A6. Visualization of data generating model for CAT-5 Math

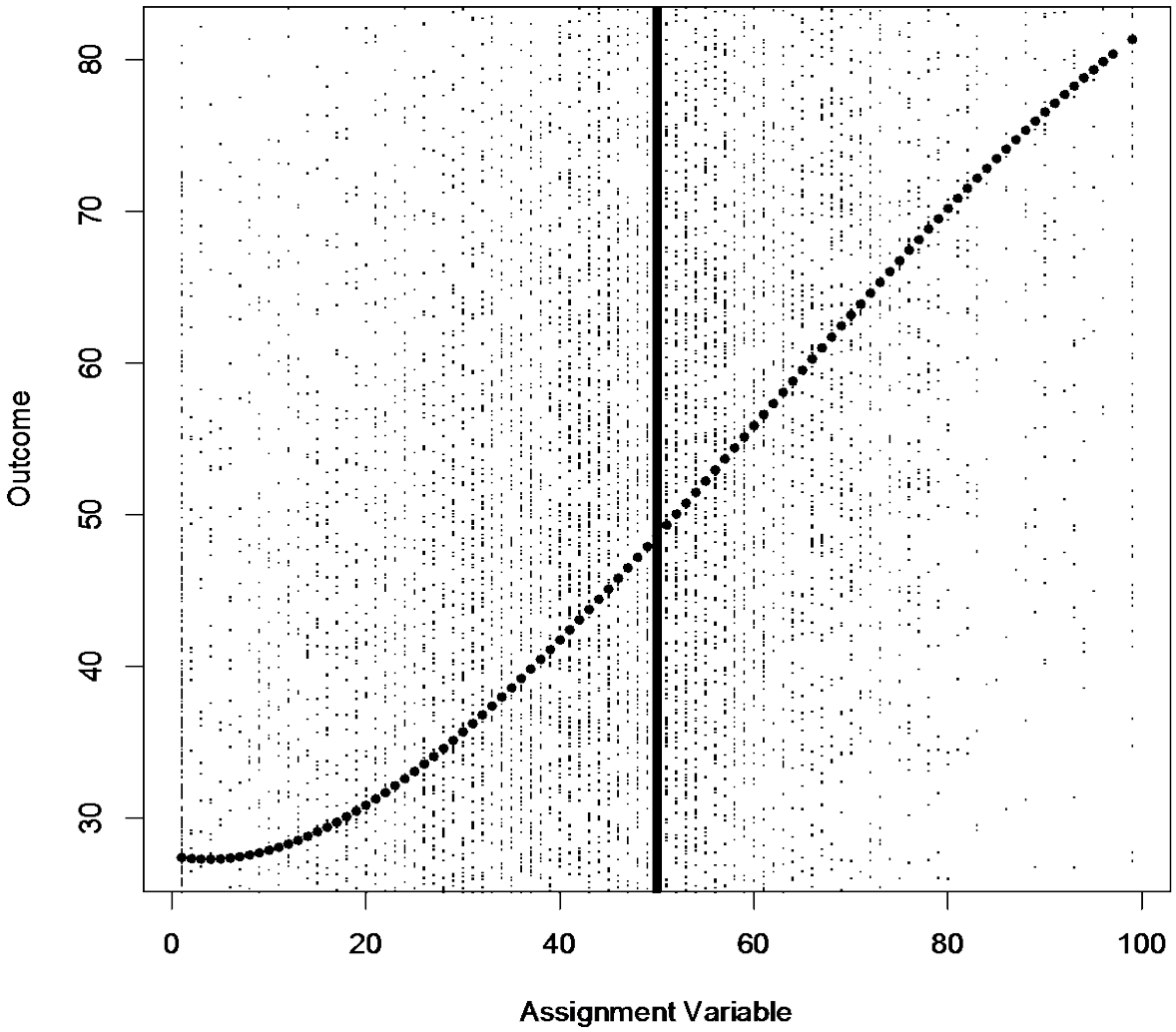


Figure A7. Visualization of data generating model for CAT-5 Reading Comprehension

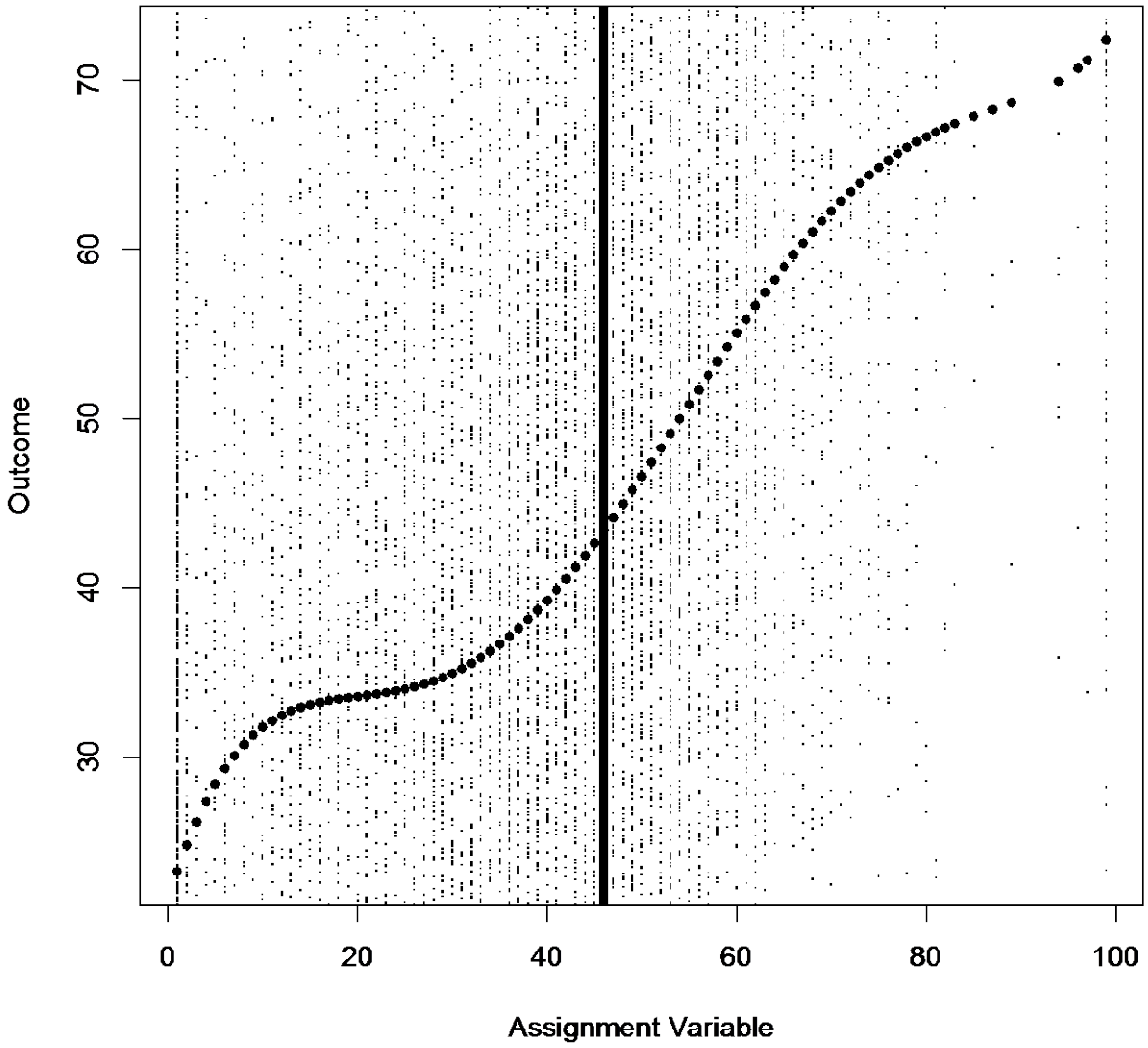


Table A10. Simulations using DGPs based on data from past evaluations: Bias

Standard Errors	Bandwidth Selection Algorithm and Sample Size					
	Imbens and Kalyanaraman (2012)			Calonico et al. (2014)		
	1,000	10,000	100,000	1,000	10,000	100,000
DGP: SAT-10 Reading (grade 1)						
Conventional	0.032	0.022	0.007	0.016	0.008	0.003
Robust	-0.005	-0.007	-0.005	0.007	-0.001	-0.002
DGP: SAT-10 Reading (grade 4)						
Conventional	-0.005	-0.003	-0.003	-0.004	-0.001	0.000
Robust	-0.004	0.002	0.000	-0.003	0.001	0.000
DGP: SAT-10 Math (grade 6)						
Conventional	0.000	-0.001	0.000	-0.001	-0.001	0.001
Robust	0.000	-0.001	0.000	-0.004	-0.003	0.001
DGP: GRADE						
Conventional	-0.035	-0.030	-0.030	-0.033	-0.030	-0.030
Robust	0.006	0.008	0.008	0.009	0.008	0.008
DGP: CAT-5 Vocabulary						
Conventional	0.005	0.004	0.002	0.005	0.004	0.002
Robust	0.001	0.000	-0.002	-0.001	-0.001	-0.002
DGP: CAT-5 Math						
Conventional	0.001	0.001	0.002	0.002	0.001	0.001
Robust	0.000	-0.001	-0.001	-0.001	-0.002	-0.001
DGP: CAT-5 Reading Comprehension						
Conventional	0.009	0.008	0.006	0.008	0.006	0.004
Robust	-0.007	-0.004	-0.002	-0.003	-0.003	-0.001

Source: Monte Carlo simulations, 10,000 replications.

Table A11. Simulations using DGPs based on data from past evaluations: Minimum Detectable Effects

Standard Errors	Bandwidth Selection Algorithm and Sample Size					
	Imbens and Kalyanaraman (2012)			Calonico et al. (2014)		
	1,000	10,000	100,000	1,000	10,000	100,000
DGP: SAT-10 Reading (grade 1)						
Conventional	0.51	0.17	0.07	0.68	0.22	0.08
Robust	0.88	0.26	0.08	0.81	0.26	0.09
DGP: SAT-10 Reading (grade 4)						
Conventional	0.43	0.14	0.04	0.52	0.16	0.06
Robust	0.77	0.24	0.07	0.62	0.19	0.09
DGP: SAT-10 Math (grade 6)						
Conventional	0.49	0.16	0.06	0.59	0.18	0.07
Robust	0.81	0.24	0.08	0.72	0.22	0.1
DGP: GRADE						
Conventional	0.47	0.15	0.05	0.48	0.15	0.05
Robust	1.01	0.28	0.09	0.83	0.28	0.09
DGP: CAT-5 Vocabulary						
Conventional	0.46	0.15	0.05	0.52	0.17	0.06
Robust	0.86	0.25	0.07	0.61	0.2	0.07
DGP: CAT-5 Math						
Conventional	0.46	0.15	0.05	0.5	0.16	0.05
Robust	0.85	0.26	0.07	0.59	0.19	0.06
DGP: CAT-5 Reading Comprehension						
Conventional	0.46	0.15	0.05	0.51	0.17	0.06
Robust	0.88	0.26	0.07	0.6	0.2	0.07

Source: Monte Carlo simulations, 10,000 replications.

Table A12. Simulations using DGPs based on data from past evaluations: Type 1 Error Rates

Standard Errors	Bandwidth Selection Algorithm and Sample Size					
	Imbens and Kalyanaraman (2012)			Calonico et al. (2014)		
	1,000	10,000	100,000	1,000	10,000	100,000
DGP: SAT-10 Reading (grade 1)						
Conventional	0.06	0.10	0.07	0.06	0.06	0.06
Robust	0.05	0.05	0.05	0.05	0.05	0.05
DGP: SAT-10 Reading (grade 4)						
Conventional	0.06	0.06	0.07	0.06	0.06	0.05
Robust	0.05	0.05	0.05	0.05	0.05	0.05
DGP: SAT-10 Math (grade 6)						
Conventional	0.06	0.05	0.05	0.06	0.06	0.05
Robust	0.05	0.05	0.05	0.05	0.05	0.05
DGP: GRADE						
Conventional	0.06	0.08	0.40	0.05	0.08	0.40
Robust	0.04	0.05	0.05	0.04	0.05	0.05
DGP: CAT-5 Vocabulary						
Conventional	0.06	0.06	0.06	0.06	0.05	0.06
Robust	0.06	0.05	0.05	0.06	0.05	0.06
DGP: CAT-5 Math						
Conventional	0.06	0.05	0.06	0.06	0.05	0.06
Robust	0.05	0.05	0.05	0.06	0.05	0.05
DGP: CAT-5 Reading Comprehension						
Conventional	0.06	0.07	0.08	0.06	0.06	0.06
Robust	0.06	0.05	0.05	0.06	0.05	0.05

Source: Monte Carlo simulations, 10,000 replications.

References

Agodini, R., and B. Harris. "An Experimental Evaluation of Four Elementary School Math Curricula." *Journal of Research on Educational Effectiveness*, vol. 3, no. 3, 2010, pp. 199–253.

Balu, R., P. Zhu, F. Doolittle, E. Schiller, J. Jenkins, and R. Gersten. "Evaluation of Response to Intervention Practices for Elementary School Reading." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2015.

Bloom, H. "Randomizing Groups to Evaluate Place-Based Programs." New York: MDRC, 2004.

Bloom, H., L. Richburg-Hayes, and A. Black. "Using Covariates to Improve Precision for Studies that Randomize Schools to Evaluate Educational Interventions." *Educational Evaluation and Policy Analysis*, vol. 29, no. 1, 2007, pp. 30–59.

Calonico, S., M. Cattaneo, and R. Titiunik. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica*, vol. 82, no. 6, 2014, pp. 2295–2326.

Campuzano, L., M. Dynarski, R. Agodini, and K. Rall. "Effectiveness of Reading and Mathematics Software Products: Findings from Two Student Cohorts." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Chiang, H., A. Wellington, K. Hallgren, C. Speroni, M. Herrmann, S. Glazerman, and J. Constantine. "Evaluation of the Teacher Incentive Fund: Implementation and Impacts of Pay-for-Performance After Two Years." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2015.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences,* 2nd edition. Mahwah, NJ: Lawrence Erlbaum Associates, 1988.

Constantine, J., D. Player, T. Silva, K. Hallgren, M. Grider, and J. Deke. "An Evaluation of Teachers Trained Through Different Routes to Certification, Final Report." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009.

Deke, J., and L. Dragoset. "Statistical Power for Regression Discontinuity Designs in Education: Empirical Estimates of Design Effects Relative to Randomized Controlled Trials." Princeton, NJ: Mathematica Policy Research, 2012.

Gelman, A., and G. Imbens. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." Cambridge, MA: National Bureau of Economic Research, 2014.

Greenberg, D., and B. S. Barnow. "Flaws in Evaluations of Social Programs: Illustrations from Randomized Controlled Trials." *Evaluation Review*, vol. 38, no. 5, 2014, pp. 359–387.

Hedges, L. V., and E. C. Hedberg. "Intraclass Correlation Values for Planning Group-Randomized Trials in Education." *Educational Evaluation and Policy Analysis,* vol. 29, no. 1, 2007, pp. 60–87.

Hill, C. J., H. S. Bloom, A. R. Black, and M. W. Lipsey. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives*, vol. 2, no. 3, 2008, pp. 172–177.

Imbens, G. W., and K. Kalyanaraman. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies*, vol. 79, no. 3, 2012, pp. 933–959.

James-Burdumy, S., J. Deke, R. Gersten, J. Lugo-Gil, R. Newman-Gonchar, J. Dimino, K. Haymond, and A. Y-H. Liu. "Effectiveness of Four Supplemental Reading Comprehension Interventions." *Journal of Research on Educational Effectiveness*, vol. 5, no. 4, 2012, pp. 345–383.

James-Burdumy, S., J. Deke, J. Lugo-Gil, N. Carey, A. Hershey, R. Gersten, R. Newman-Gonchar, J. Dimino, K. Haymond, and B. Faddis. "Effectiveness of Selected Supplemental Reading Comprehension Interventions: Findings from Two Student Cohorts." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2010.

James-Burdumy, S., M. Dynarski, and J. Deke. "After-School Program Effects on Behavior: Results from the 21st Century Community Learning Centers Program National Evaluation." *Economic Inquiry*, vol. 46, no. 1, 2008, pp. 13–18.

Kane, T. J. "Frustrated with the Pace of Progress in Education? Invest in Better Evidence." Washington, DC: The Brookings Institution, 2015.

Leamer, E. "Tantalus on the Road to Asymptopia." *Journal of Economic Perspectives*, vol. 24, no. 2, 2010, pp. 31–46.

Lipsey, M. W., K. Puzio, C. Yun, M. A. Hebert, K. Steinka-Fry, M. W. Cole, M. Roberts, K. S. Anthony, and M. D. Busick. "Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms." Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education, 2012.

Murray, D. M. *Design and Analysis of Group-Randomized Trials*. Oxford, United Kingdom: Oxford University Press, 1998.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Schochet, P. Z. "Statistical Power for Random Assignment Evaluations of Education Programs." *Journal of Educational and Behavioral Statistics,* vol. 33, no. 1, 2008a, pp. 62–87.

Schochet, P. Z. "Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluations." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2008b.

What Works Clearinghouse. "WWC Procedures and Standards Handbook (Version 2.0)." Washington, DC: Institute of Education Sciences, U.S. Department of Education, 2008.

What Works Clearinghouse. "Assessing Attrition Bias (Version 2.1)." Washington, DC: Institute of Education Sciences, U.S. Department of Education, 2013.

What Works Clearinghouse. "Assessing Attrition Bias—Addendum (Version 3.0)." Washington, DC: Institute of Education Sciences, U.S. Department of Education, 2014.